Logzip: Extracting Hidden Structures via Iterative Clustering for Log Compression

Jinyang Liu[1], Jieming Zhu[2], Shilin He[3], Pinjia He[4], Zibin Zheng[1], **Michael R. Lyu**[3]

[1]Sun Yat-Sen University

[2]Huawei Noah's Ark Lab

[3]The Chinese University of Hong Kong

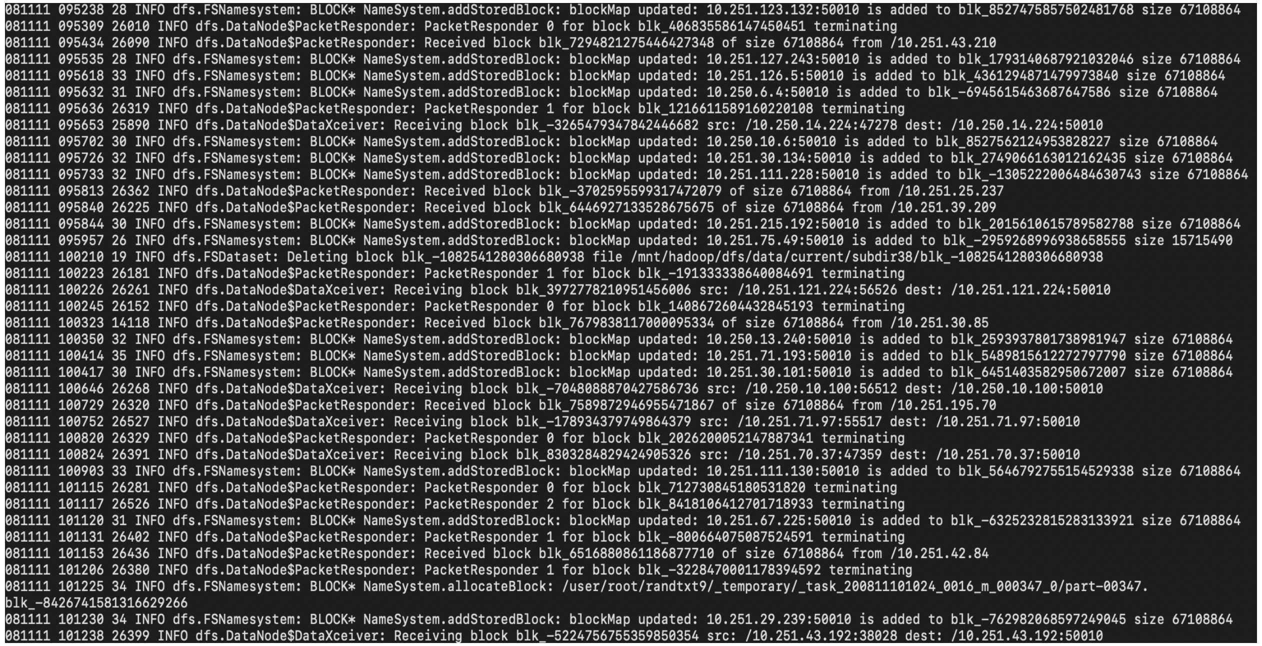
[4]ETH Zurich

Supervisor: Prof. Zibin Zheng and Prof. Michael R. Lyu

The 34th IEEE/ACM International Conference on Automated Software Engineering (ASE)

Motivation & Background

Systems produce logs to record runtime information



```
081111 095238 28 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.123.132:50010 is added to blk_8527475857502481768 size 67108864
081111 095309 26010 INFO dfs.DataNode$PacketResponder: PacketResponder 0 for block blk_4068355861474504 51 terminating
081111 095434 26090 INFO dfs.DataNode$PacketResponder: Received block blk_7294821275446427348 of size 67108864 from /10.251.43.210
081111 095535 28 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.127.243:50010 is added to blk_1793140687921032046 size 67108864
081111 095618 33 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.126.5:50010 is added to blk_4361294871479973840 size 67108864
081111 095632 31 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.250.6.4:50010 is added to blk_-6945615463687647586 size 67108864
081111 095636 26319 INFO dfs.DataNode$PacketResponder: PacketResponder 1 for block blk_1216611589160220108 terminating
081111 095653 25890 INFO dfs.DataNode$DataXceiver: Receiving block blk_-3265479347842446682 src: /10.250.14.224:47278 dest: /10.250.14.224:50010
081111 095702 30 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.250.10.6:50010 is added to blk_852756212495382822 7 size 67108864
081111 095726 32 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.30.134:50010 is added to blk_2749066163012162435 size 67108864
081111 095733 32 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.111.228:50010 is added to blk_-1305222006484630743 size 67108864
081111 095813 26362 INFO dfs.DataNode$PacketResponder: Received block blk_-3702595599317472079 of size 67108864 from /10.251.25.237
081111 095840 26225 INFO dfs.DataNode$PacketResponder: Received block blk_6446927133528675675 of size 67108864 from /10.251.39.209
081111 095844 30 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.215.192:50010 is added to blk_2015610615789582788 size 67108864
081111 095957 26 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.75.49:50010 is added to blk_-2959268996938658555 size 15715490
081111 100210 19 INFO dfs.FSDataset: Deleting block blk_-1082541280306680938 file /mnt/hadoop/dfs/data/current/subdir38/blk_-1082541280306680938
081111 100223 26181 INFO dfs.DataNode$PacketResponder: PacketResponder 1 for block blk_-191333338640084691 terminating
081111 100226 26261 INFO dfs.DataNode$DataXceiver: Receiving block blk_3972778210951456006 src: /10.251.121.224:56526 dest: /10.251.121.224:50010
081111 100245 26152 INFO dfs.DataNode$PacketResponder: PacketResponder 0 for block blk_1408672604432845193 terminating
081111 100323 14118 INFO dfs.DataNode$PacketResponder: Received block blk_7679838117000095334 of size 67108864 from /10.251.30.85
081111 100350 32 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.250.13.240:50010 is added to blk_2593937801738981947 size 67108864
081111 100414 35 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.71.193:50010 is added to blk_5489815612272797790 size 67108864
081111 100417 30 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.30.101:50010 is added to blk_6451403582950672007 size 67108864
081111 100646 26268 INFO dfs.DataNode$DataXceiver: Receiving block blk_-7048088870427586736 src: /10.250.10.100:56512 dest: /10.250.10.100:50010
081111 100729 26320 INFO dfs.DataNode$PacketResponder: Received block blk_7589872946955471867 of size 67108864 from /10.251.195.70
081111 100752 26527 INFO dfs.DataNode$DataXceiver: Receiving block blk_-178934379749864379 src: /10.251.71.97:55517 dest: /10.251.71.97:50010
081111 100820 26329 INFO dfs.DataNode$PacketResponder: PacketResponder 0 for block blk_2026200052147887341 terminating
081111 100824 26391 INFO dfs.DataNode$DataXceiver: Receiving block blk_8303284829424905326 src: /10.251.70.37:47359 dest: /10.251.70.37:50010
081111 100903 33 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.111.130:50010 is added to blk_5646792755154529338 size 67108864
081111 101115 26281 INFO dfs.DataNode$PacketResponder: PacketResponder 0 for block blk_712730845180531820 terminating
081111 101117 26526 INFO dfs.DataNode$PacketResponder: PacketResponder 2 for block blk_8418106412701718933 terminating
081111 101120 31 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.67.225:50010 is added to blk_-6325232815283133921 size 67108864
081111 101131 26402 INFO dfs.DataNode$PacketResponder: PacketResponder 1 for block blk_-800664075087524591 terminating
081111 101153 26436 INFO dfs.DataNode$PacketResponder: Received block blk_6516880861186877710 of size 67108864 from /10.251.42.84
081111 101206 26380 INFO dfs.DataNode$PacketResponder: PacketResponder 1 for block blk_-3228470001178394592 terminating
081111 101225 34 INFO dfs.FSNamesystem: BLOCK* NameSystem.allocateBlock: /user/root/randtxt9/_temporary/_task_200811101024_0016_m_000347_0/part-00347.
blk_-8426741581316629266
081111 101230 34 INFO dfs.FSNamesystem: BLOCK* NameSystem.addStoredBlock: blockMap updated: 10.251.29.239:50010 is added to blk_-762982068597249045 size 67108864
081111 101238 26399 INFO dfs.DataNode$DataXceiver: Receiving block blk_-5224756755359850354 src: /10.251.43.192:38028 dest: /10.251.43.192:50010
```

Motivation & Background

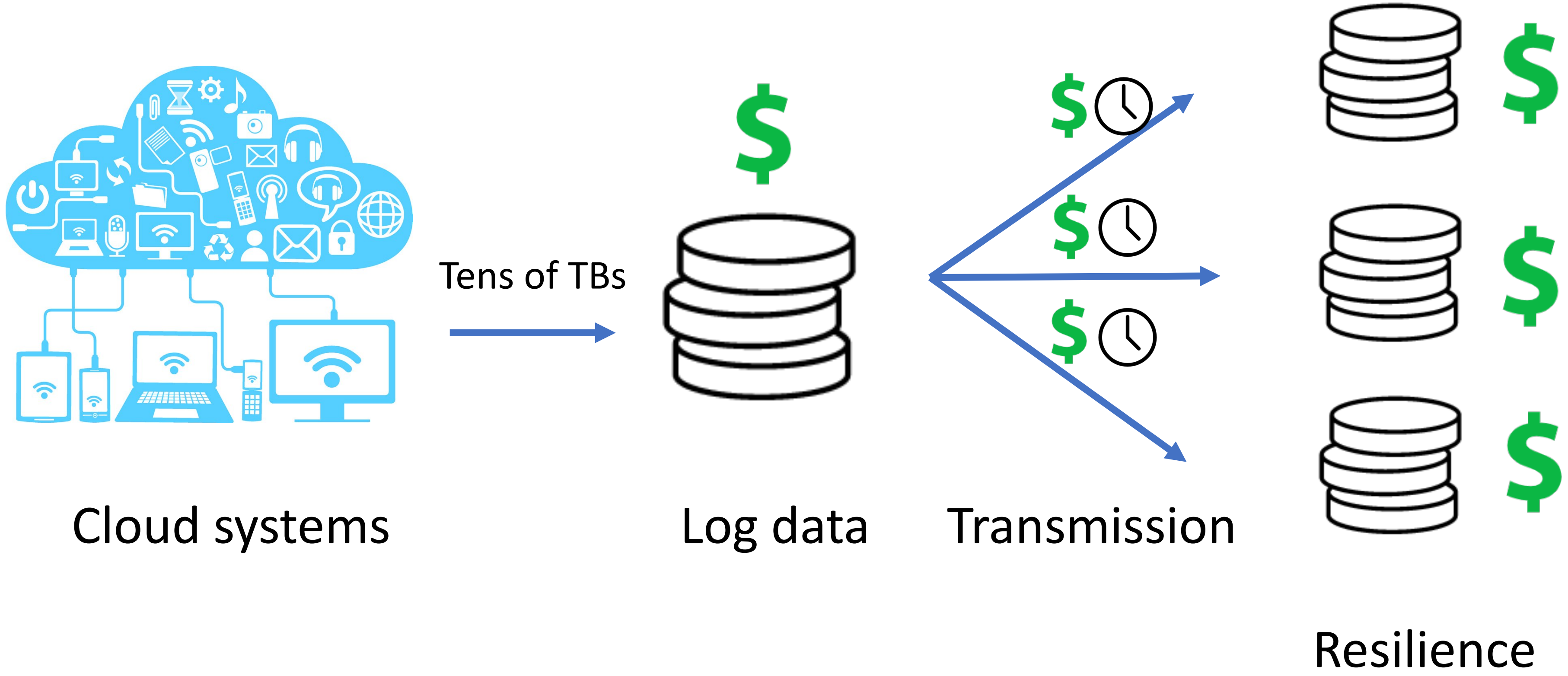
System logs are important for

- Diagnose runtime failures
- Identify performance bottlenecks
- Detect security issues
- Market trends prediction

……

Motivation & Background

Log data requires long-term storage and is fast-growing



Tens of TBs

Cloud systems

Log data

Transmission

Resilience

It is **time**-comsuming and **money**-consuming

Motivation & Background

How to reduce data storage cost?

Writing less logging statements in the source code

↓

Risk missing key information ☹

Apply compression tools: gzip, bzip2, lzma...

↓

Not specifically designed for log data ☹

Logzip, explores **hidden structures** of log data for better compression ☺

Motivation & Background

Intuition: repetitive data is more compressible

The Chinese University of Hong Kong is wonderful!
The Chinese University of Hong Kong is wonderful!
The Chinese University of Hong Kong is wonderful!
The Chinese University of Hong Kong is wonderful!
The Chinese University of Hong Kong is wonderful!
The Chinese University of Hong Kong is wonderful!

(~300 chars)

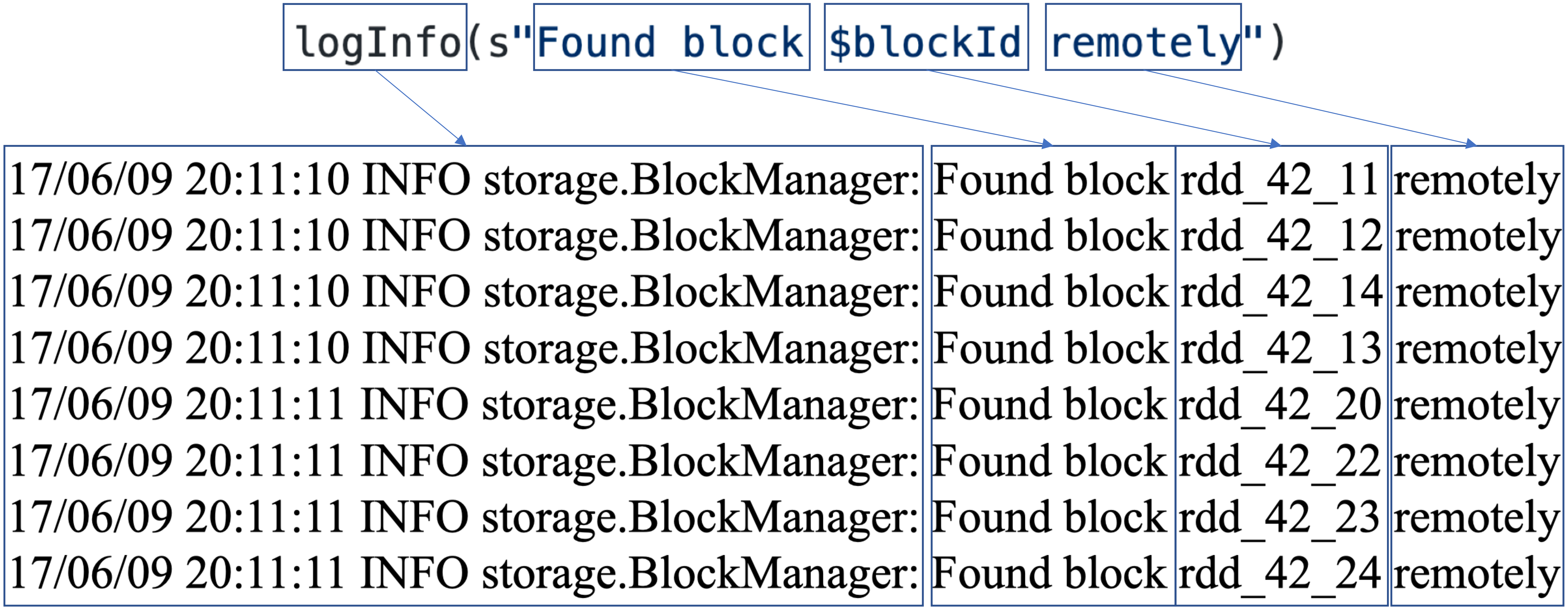
The Chinese University of Hong Kong (CUHK) is a public research university in Shatin, Hong Kong formally established in 1963 by a charter granted by the Legislative Council of Hong Kong. It is the territory's second oldest university and was founded as a federation of three existing colleges.

(~300 chars)

Easy to handle!
I can somewhat handle it.
well
gzip

Motivation & Background

Log Structure



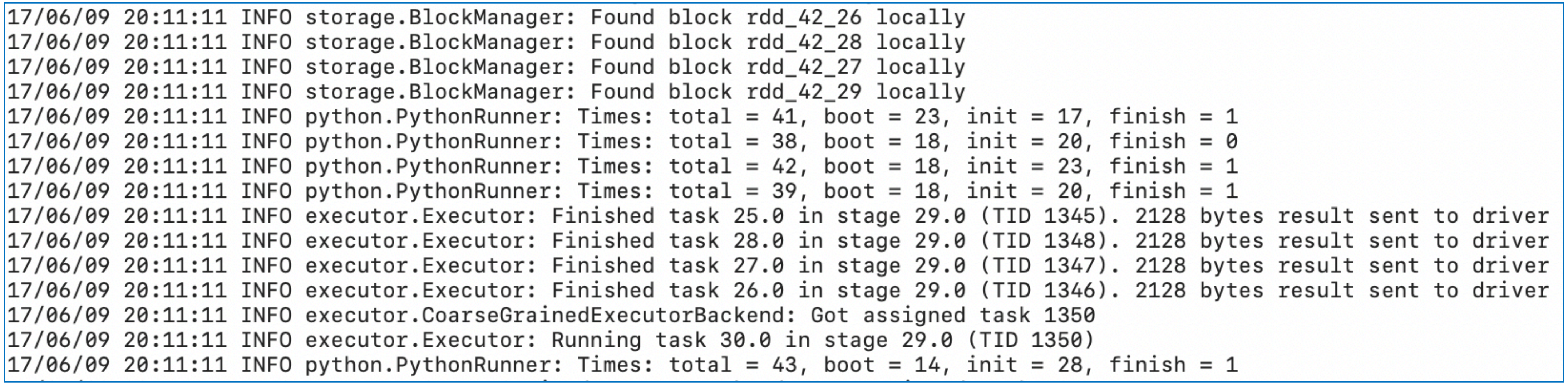
```
logInfo(s"Found block $blockId remotely")
```

```
17/06/09 20:11:10 INFO storage.BlockManager: Found block rdd_42_11 remotely
17/06/09 20:11:10 INFO storage.BlockManager: Found block rdd_42_12 remotely
17/06/09 20:11:10 INFO storage.BlockManager: Found block rdd_42_14 remotely
17/06/09 20:11:10 INFO storage.BlockManager: Found block rdd_42_13 remotely
17/06/09 20:11:11 INFO storage.BlockManager: Found block rdd_42_20 remotely
17/06/09 20:11:11 INFO storage.BlockManager: Found block rdd_42_22 remotely
17/06/09 20:11:11 INFO storage.BlockManager: Found block rdd_42_23 remotely
17/06/09 20:11:11 INFO storage.BlockManager: Found block rdd_42_24 remotely
```

HDFS logs (part)

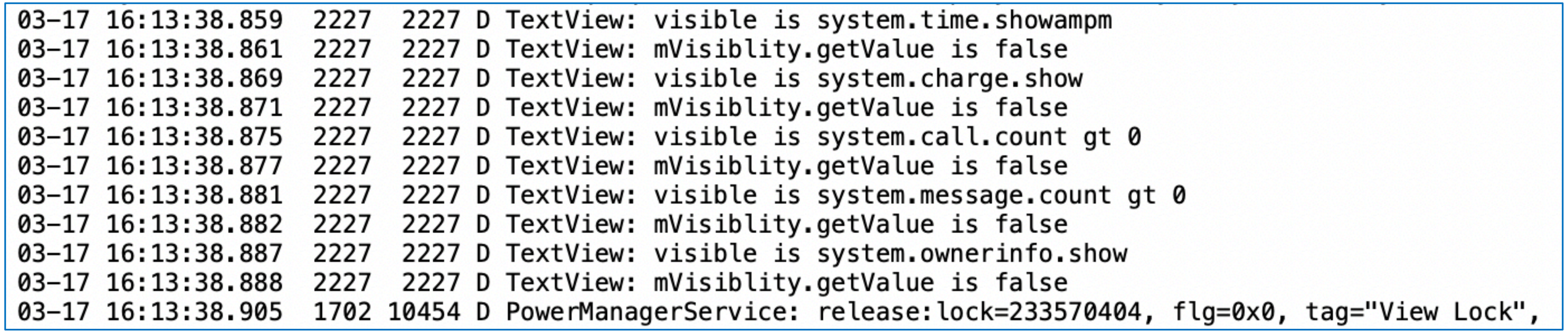
Motivation & Background

Different types of log data share the similar format



```
17/06/09 20:11:11 INFO storage.BlockManager: Found block rdd_42_26 locally
17/06/09 20:11:11 INFO storage.BlockManager: Found block rdd_42_28 locally
17/06/09 20:11:11 INFO storage.BlockManager: Found block rdd_42_27 locally
17/06/09 20:11:11 INFO storage.BlockManager: Found block rdd_42_29 locally
17/06/09 20:11:11 INFO python.PythonRunner: Times: total = 41, boot = 23, init = 17, finish = 1
17/06/09 20:11:11 INFO python.PythonRunner: Times: total = 38, boot = 18, init = 20, finish = 0
17/06/09 20:11:11 INFO python.PythonRunner: Times: total = 42, boot = 18, init = 23, finish = 1
17/06/09 20:11:11 INFO python.PythonRunner: Times: total = 39, boot = 18, init = 20, finish = 1
17/06/09 20:11:11 INFO executor.Executor: Finished task 25.0 in stage 29.0 (TID 1345). 2128 bytes result sent to driver
17/06/09 20:11:11 INFO executor.Executor: Finished task 28.0 in stage 29.0 (TID 1348). 2128 bytes result sent to driver
17/06/09 20:11:11 INFO executor.Executor: Finished task 27.0 in stage 29.0 (TID 1347). 2128 bytes result sent to driver
17/06/09 20:11:11 INFO executor.Executor: Finished task 26.0 in stage 29.0 (TID 1346). 2128 bytes result sent to driver
17/06/09 20:11:11 INFO executor.CoarseGrainedExecutorBackend: Got assigned task 1350
17/06/09 20:11:11 INFO executor.Executor: Running task 30.0 in stage 29.0 (TID 1350)
17/06/09 20:11:11 INFO python.PythonRunner: Times: total = 43, boot = 14, init = 28, finish = 1
```

Spark logs (part)

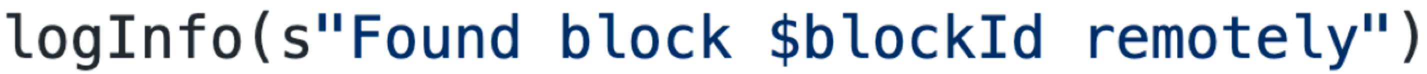


```
03-17 16:13:38.859  2227  2227 D TextView: visible is system.time.showampm
03-17 16:13:38.861  2227  2227 D TextView: mVisiblity.getValue is false
03-17 16:13:38.869  2227  2227 D TextView: visible is system.charge.show
03-17 16:13:38.871  2227  2227 D TextView: mVisiblity.getValue is false
03-17 16:13:38.875  2227  2227 D TextView: visible is system.call.count gt 0
03-17 16:13:38.877  2227  2227 D TextView: mVisiblity.getValue is false
03-17 16:13:38.881  2227  2227 D TextView: visible is system.message.count gt 0
03-17 16:13:38.882  2227  2227 D TextView: mVisiblity.getValue is false
03-17 16:13:38.887  2227  2227 D TextView: visible is system.ownerinfo.show
03-17 16:13:38.888  2227  2227 D TextView: mVisiblity.getValue is false
03-17 16:13:38.905  1702 10454 D PowerManagerService: release:lock=233570404, flg=0x0, tag="View Lock",
```

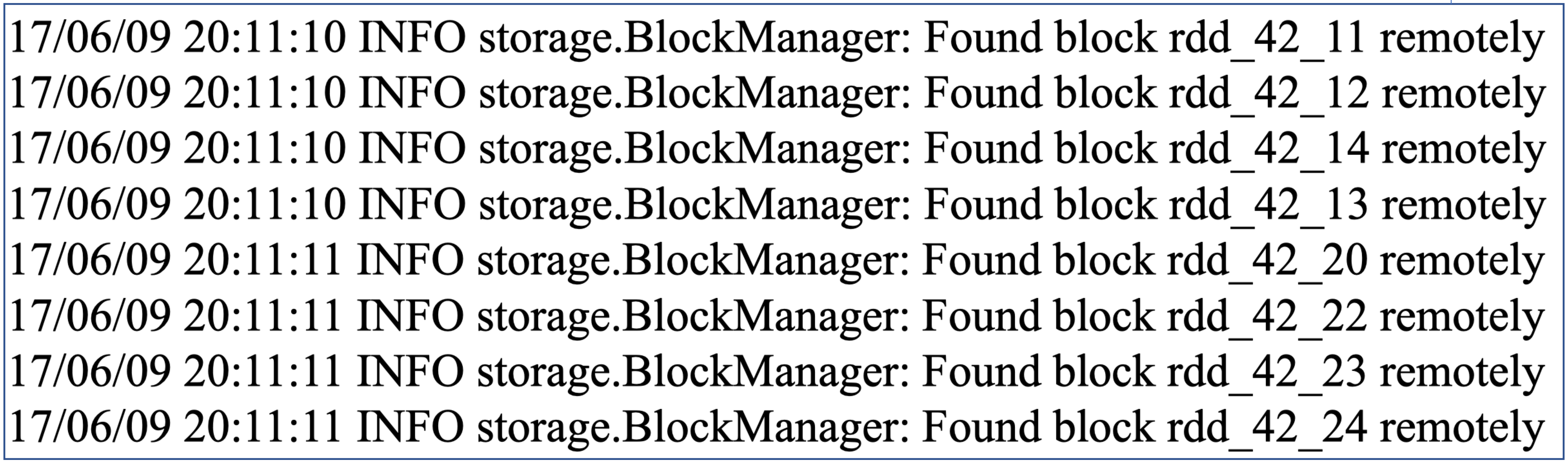
Android logs (part)

Motivation & Background

It is OK to compress the whole file



```
logInfo(s"Found block $blockId remotely")
```

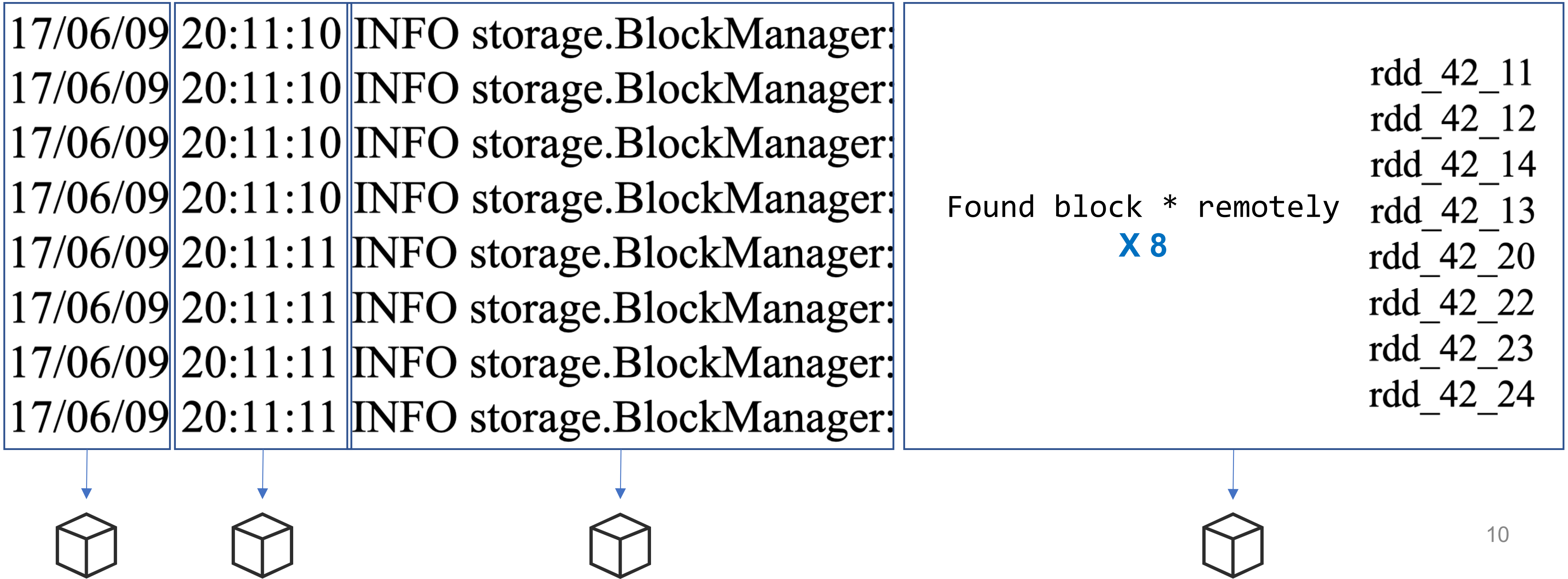


```
17/06/09 20:11:10 INFO storage.BlockManager: Found block rdd_42_11 remotely
17/06/09 20:11:10 INFO storage.BlockManager: Found block rdd_42_12 remotely
17/06/09 20:11:10 INFO storage.BlockManager: Found block rdd_42_14 remotely
17/06/09 20:11:10 INFO storage.BlockManager: Found block rdd_42_13 remotely
17/06/09 20:11:11 INFO storage.BlockManager: Found block rdd_42_20 remotely
17/06/09 20:11:11 INFO storage.BlockManager: Found block rdd_42_22 remotely
17/06/09 20:11:11 INFO storage.BlockManager: Found block rdd_42_23 remotely
17/06/09 20:11:11 INFO storage.BlockManager: Found block rdd_42_24 remotely
```

Motivation & Background

It is better to compress after hidden structures are extracted

```
logInfo(s"Found block $blockId remotely")
```



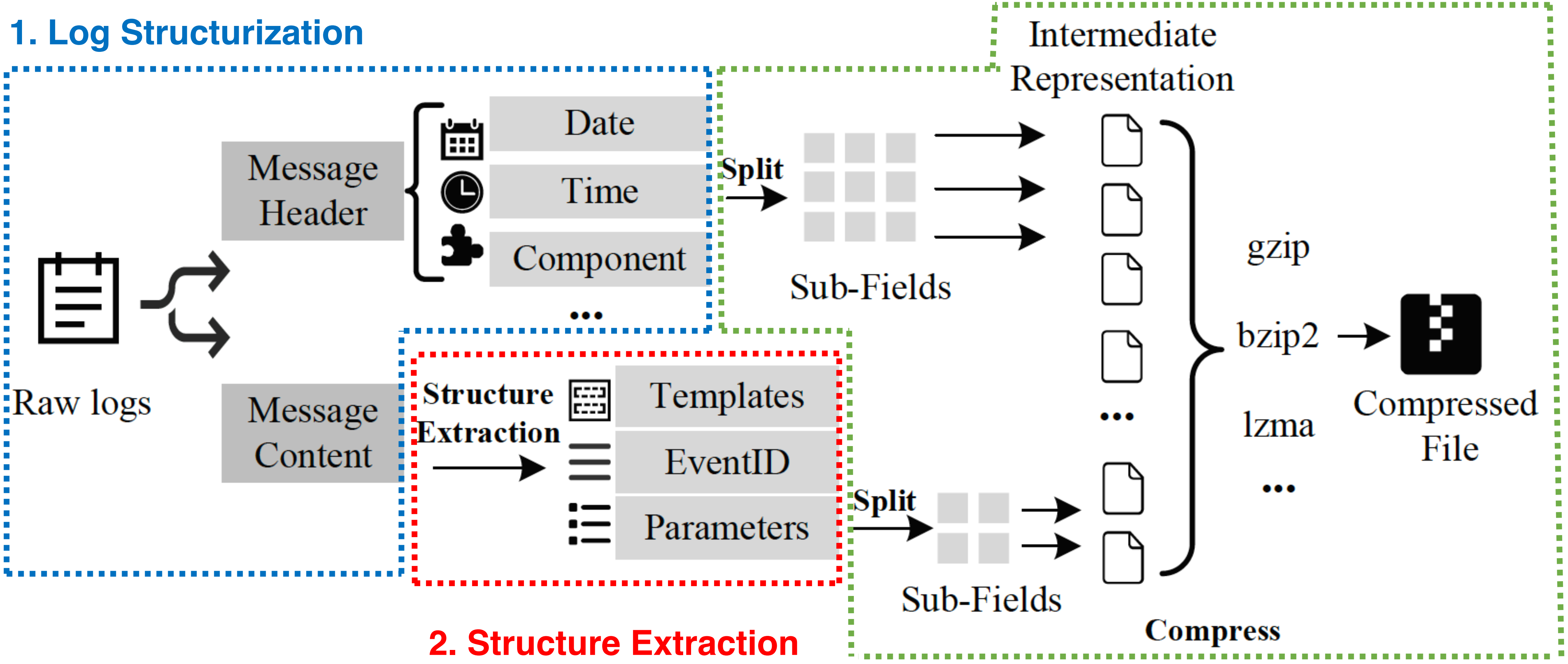
```
17/06/09 20:11:10 INFO storage.BlockManager:
17/06/09 20:11:10 INFO storage.BlockManager:
17/06/09 20:11:10 INFO storage.BlockManager:
17/06/09 20:11:10 INFO storage.BlockManager:
17/06/09 20:11:11 INFO storage.BlockManager:
17/06/09 20:11:11 INFO storage.BlockManager:
17/06/09 20:11:11 INFO storage.BlockManager:
17/06/09 20:11:11 INFO storage.BlockManager:
```

`Found block * remotely`

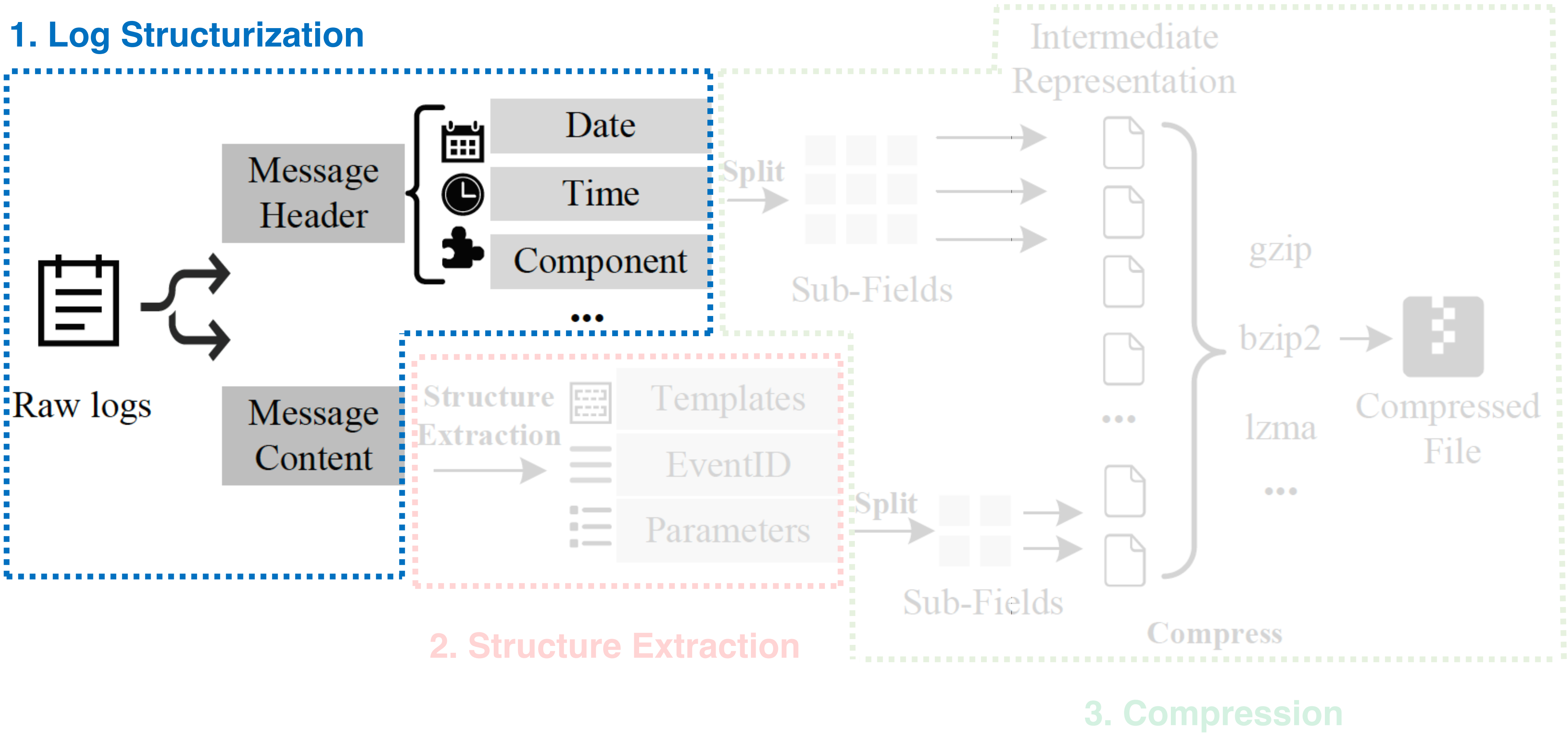
X 8

```
rdd_42_11
rdd_42_12
rdd_42_14
rdd_42_13
rdd_42_20
rdd_42_22
rdd_42_23
rdd_42_24
```

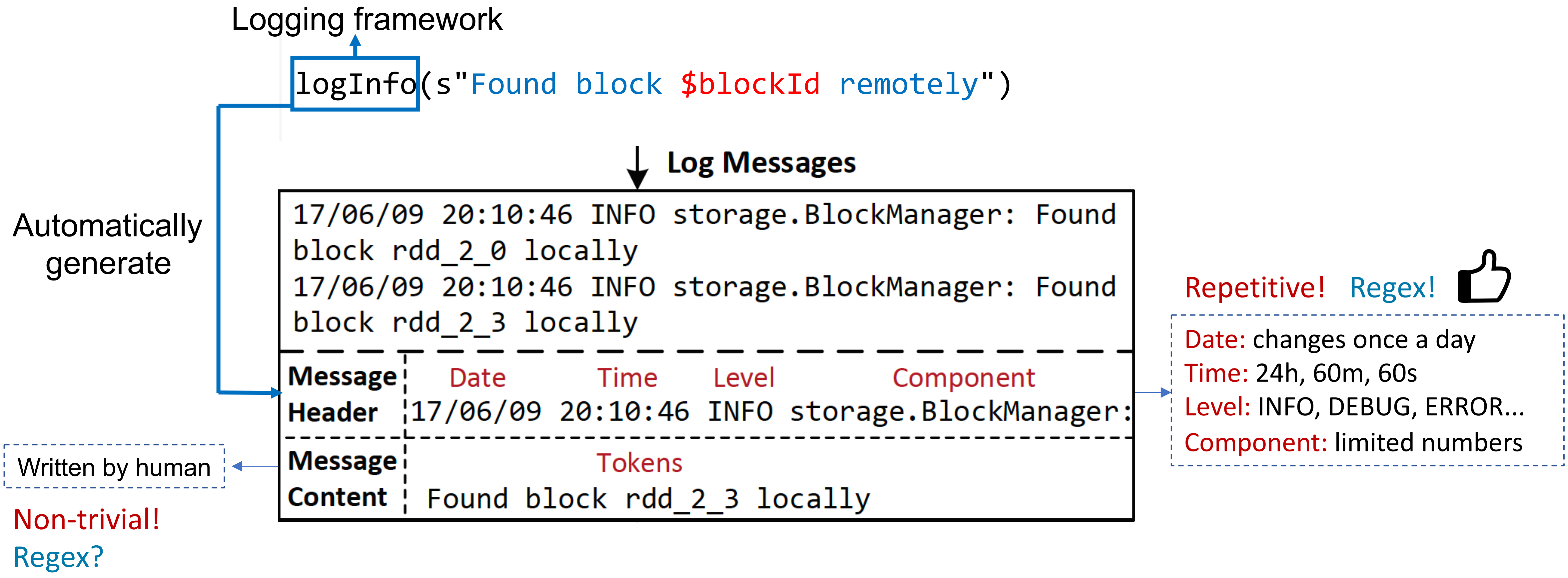
Method: Overview



Method: Log Structurization



Method: Log Structurization



Logging framework

```
logInfo(s"Found block $blockId remotely")
```

Automatically generate

Log Messages

```
17/06/09 20:10:46 INFO storage.BlockManager: Found
block rdd_2_0 locally
17/06/09 20:10:46 INFO storage.BlockManager: Found
block rdd_2_3 locally
```

Message Header

```
 Date      Time     Level   Component
17/06/09 20:10:46 INFO storage.BlockManager:
```

Message Content

```
            Tokens
Found block rdd_2_3 locally
```

Written by human

Non-trivial!
Regex?

Repetitive! Regex!

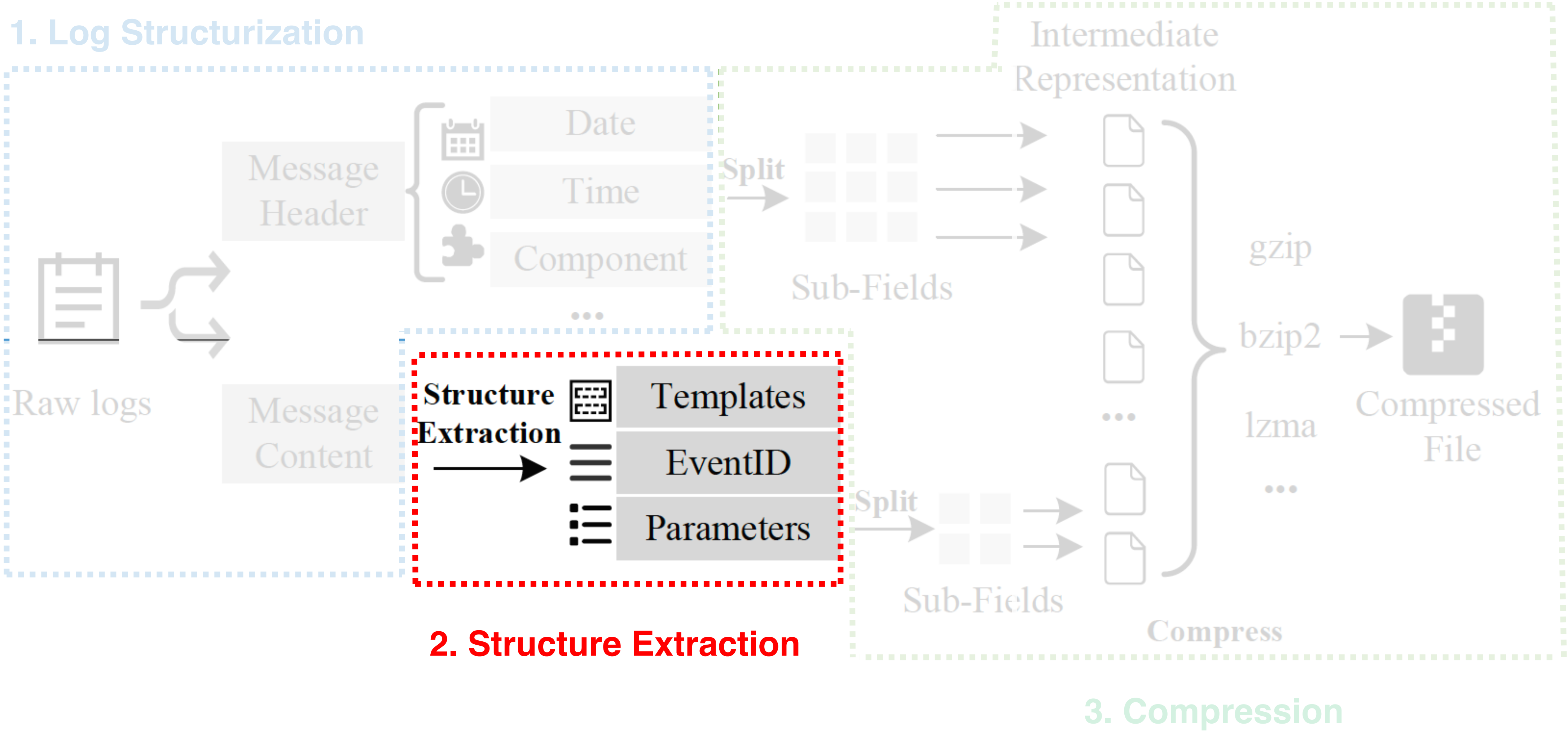
Date: changes once a day
Time: 24h, 60m, 60s
Level: INFO, DEBUG, ERROR…
Component: limited numbers

We p … cture

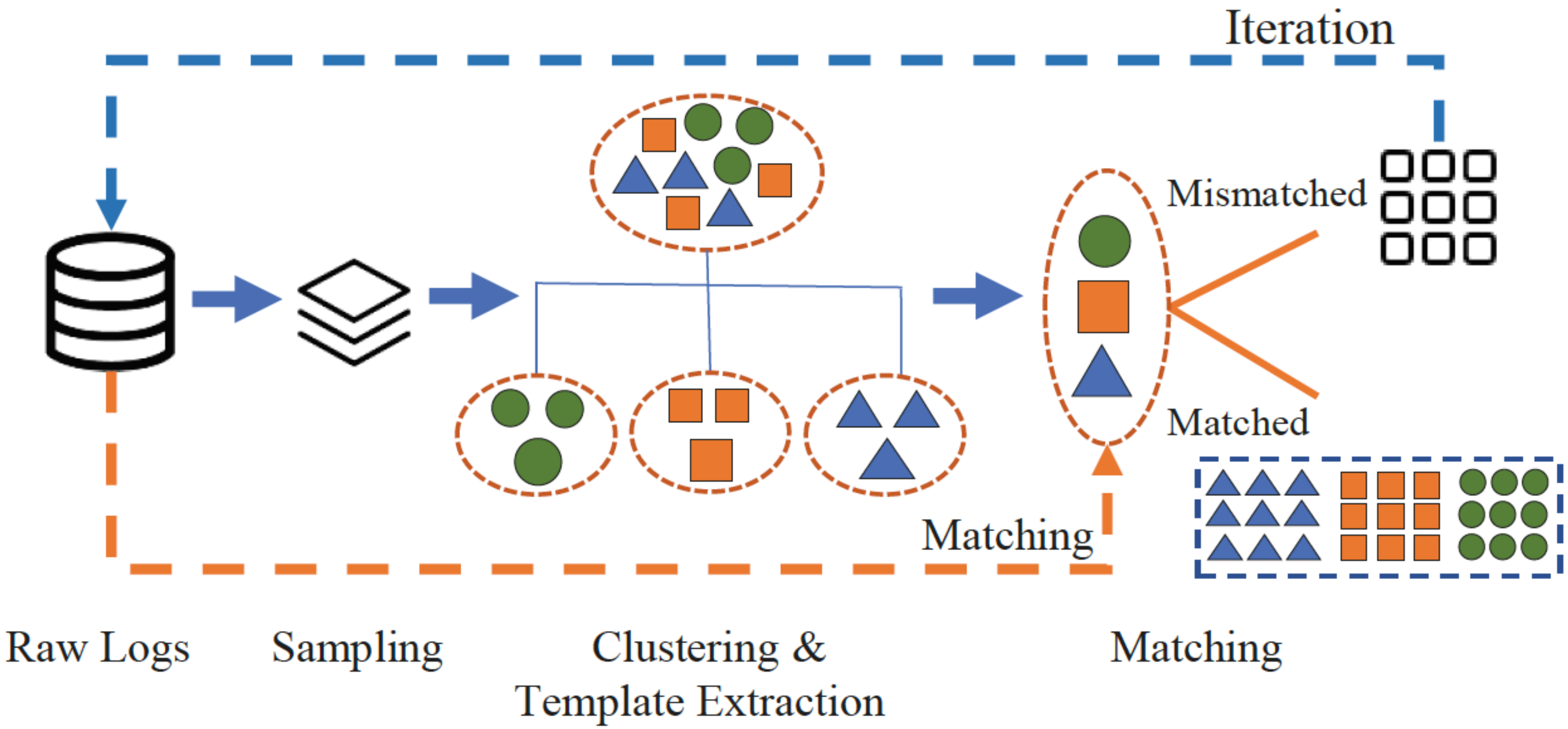
Sample regex for Android

```
[r'(/[\w-]+)+', r'([\w-]+\.){2,}[\w-]+', r'\b(\-?\+?\d+)\b|\b0[Xx][a-fA-F\d]+\b|\b[a-fA-F\d]{4,}\b']
```

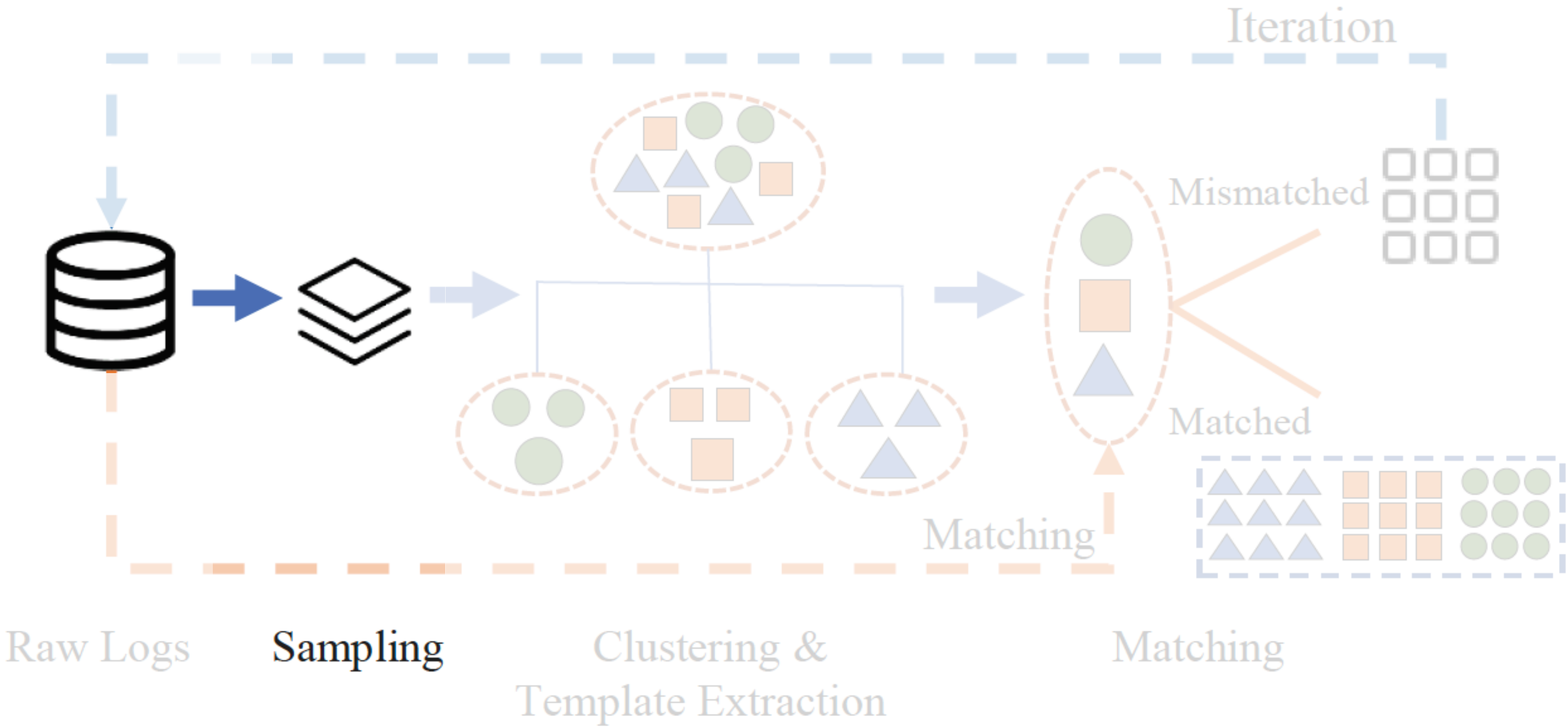
Method: Structure Extraction



Method: Iterative Structure Extraction (ISE)

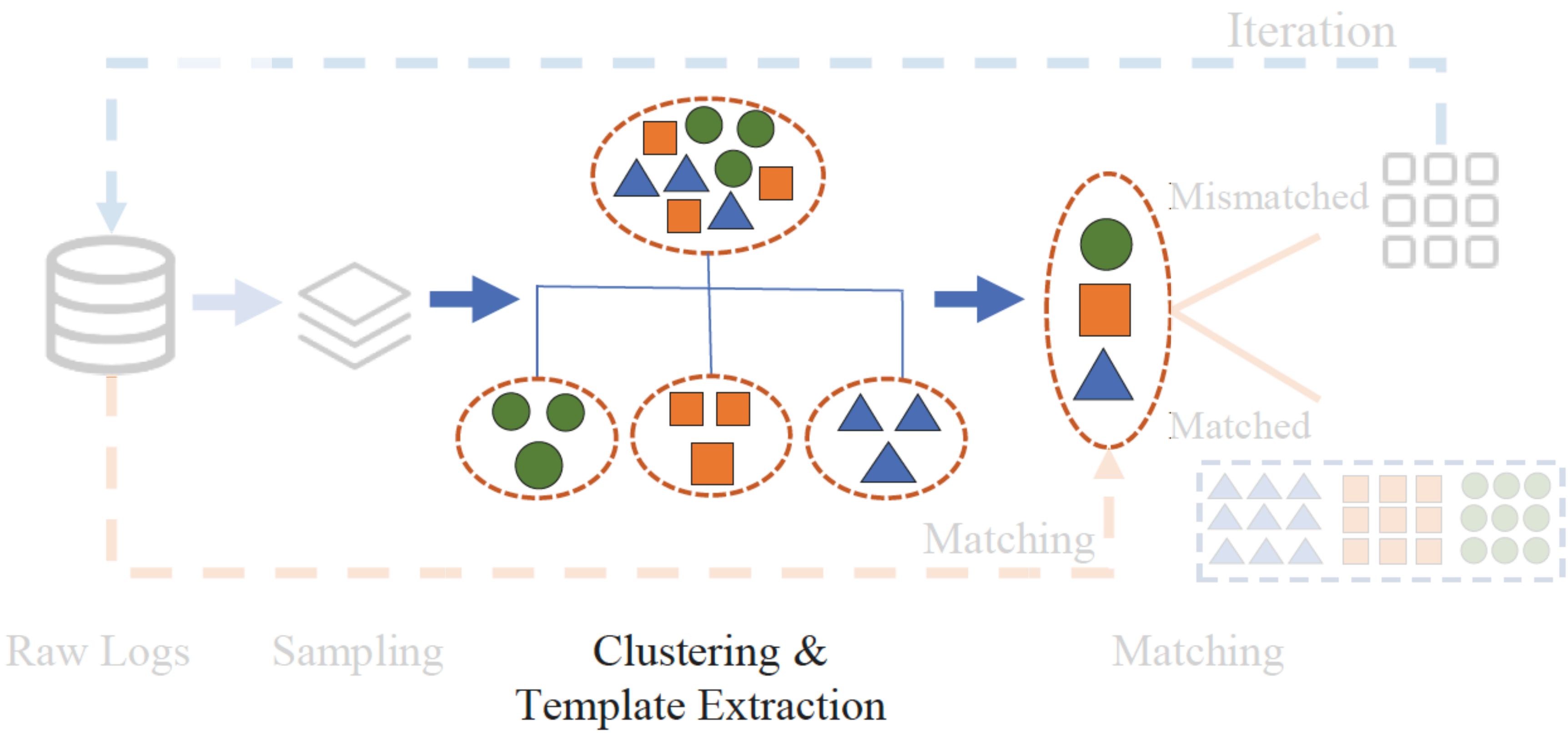


Method: Iterative Structure Extraction (ISE)



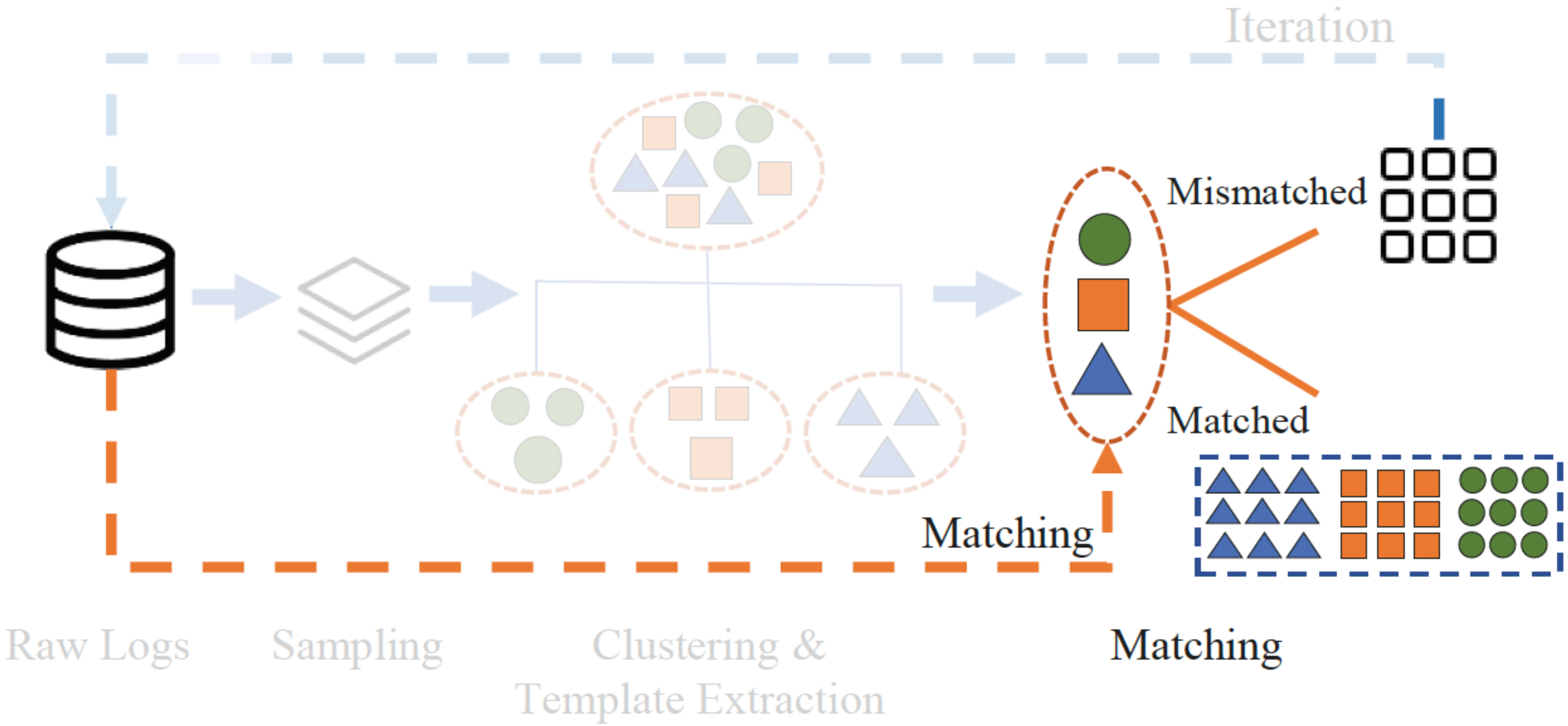
Sample a small fraction (1%) of the whole dataset

Method: Iterative Structure Extraction (ISE)



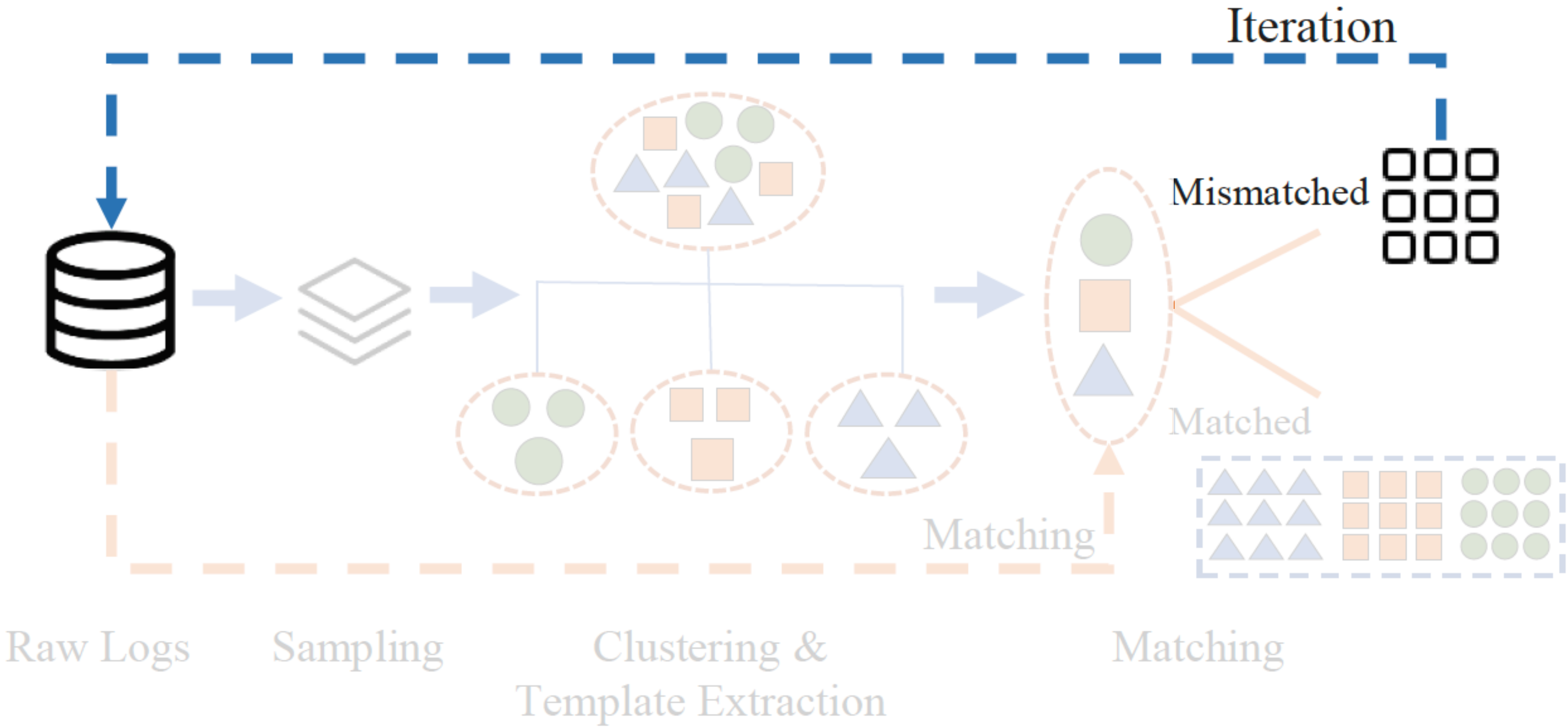
Apply sequential clustering to extract templates

Method: Iterative Structure Extraction (ISE)



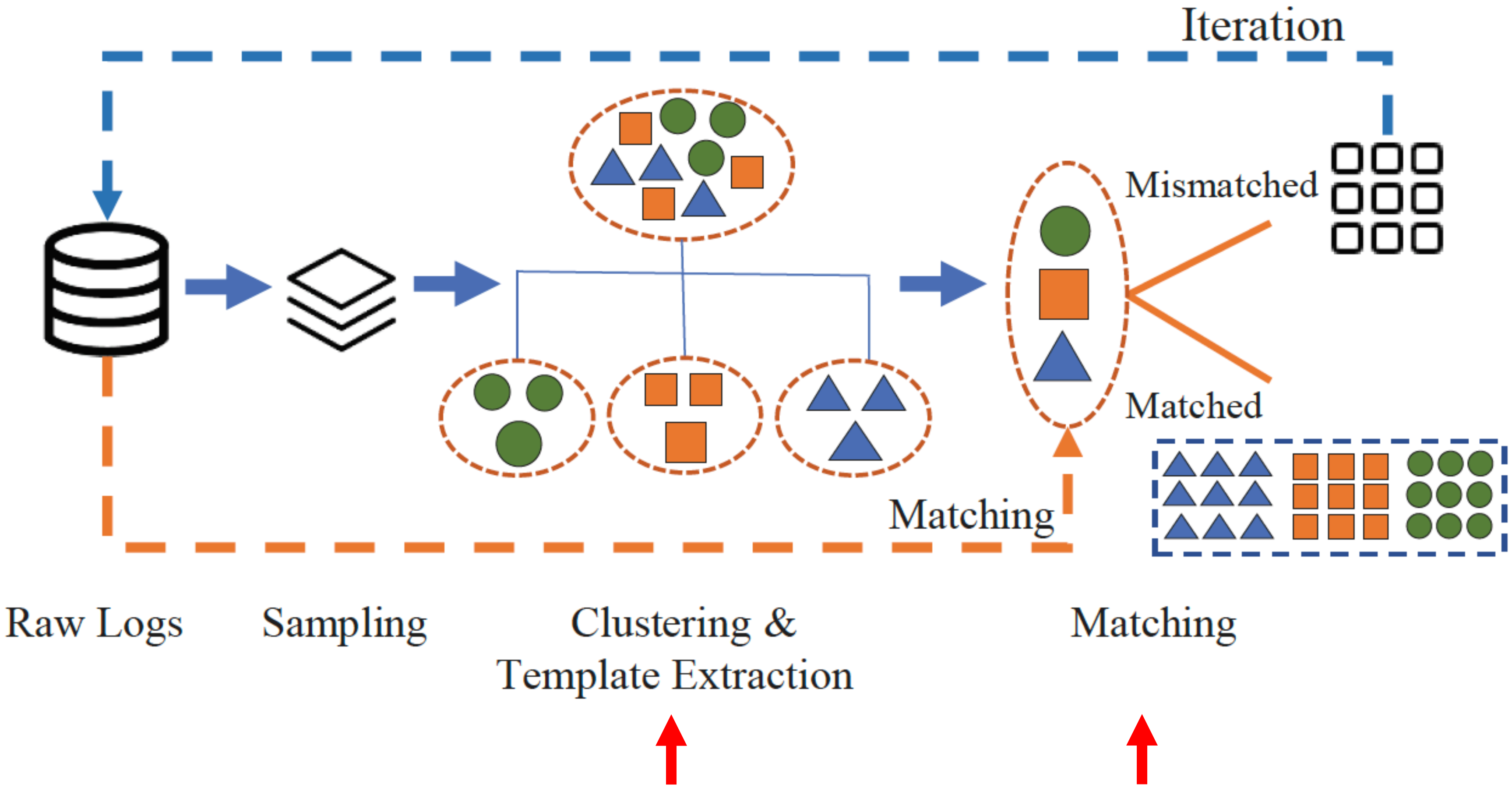
Match unsampled data with extracted templates

Method: Iterative Structure Extraction (ISE)

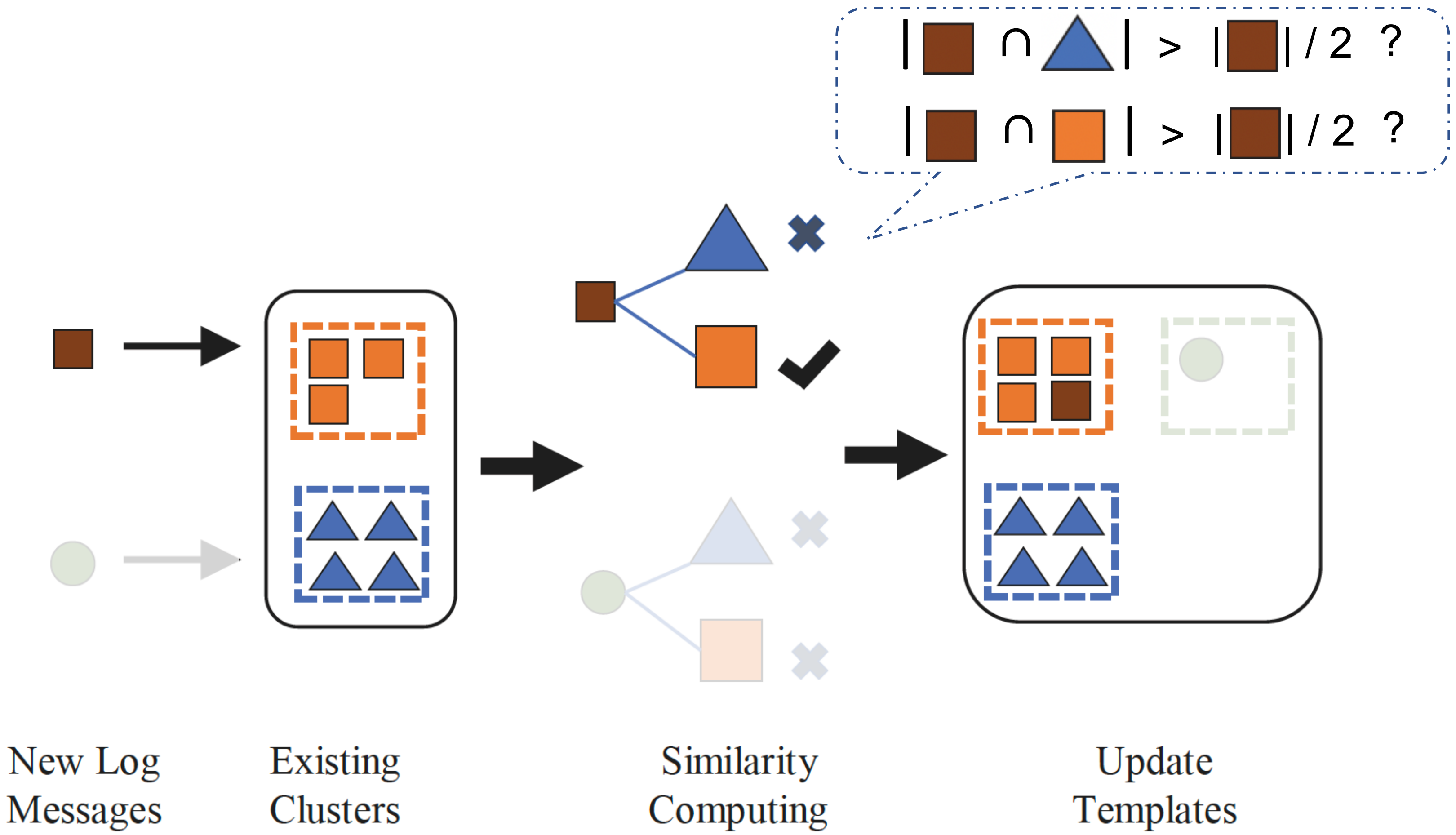


Mismatched data goes through the process iteratively

Method: Iterative Structure Extraction (ISE)



Method: ISE-Clustering & Template Extraction



New Log Messages

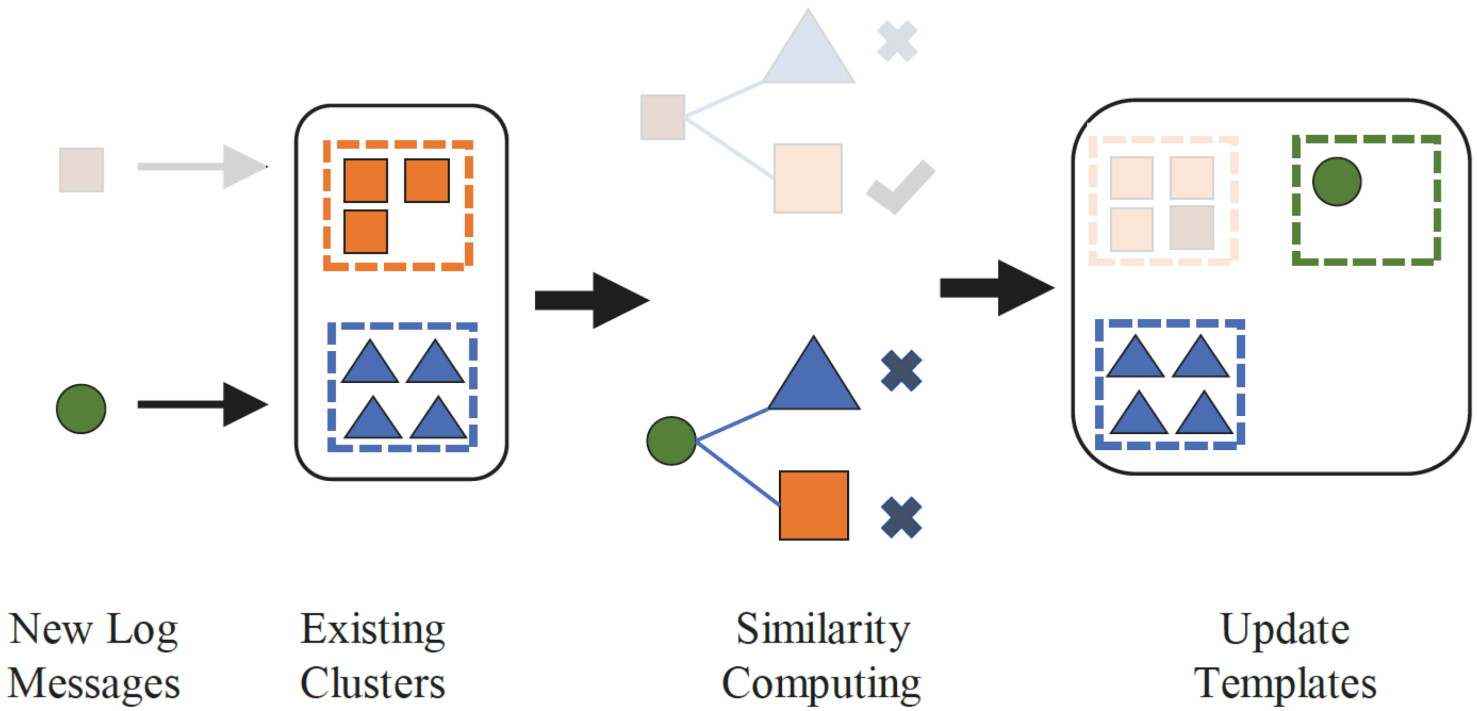
Existing Clusters

Similarity Computing

Update Templates

Workflow of Sequential Clustering

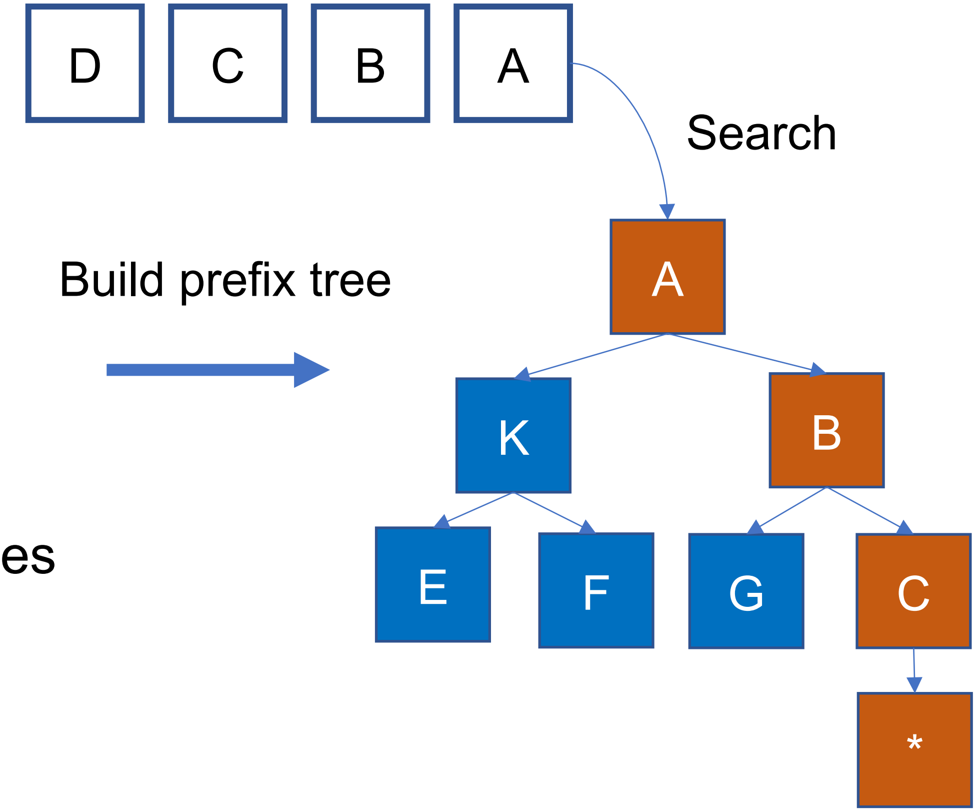
Method: ISE-Clustering & Template Extraction



Workflow of Sequential Clustering

Method: ISE-Matching

Input 1: A tokenized log message



Search

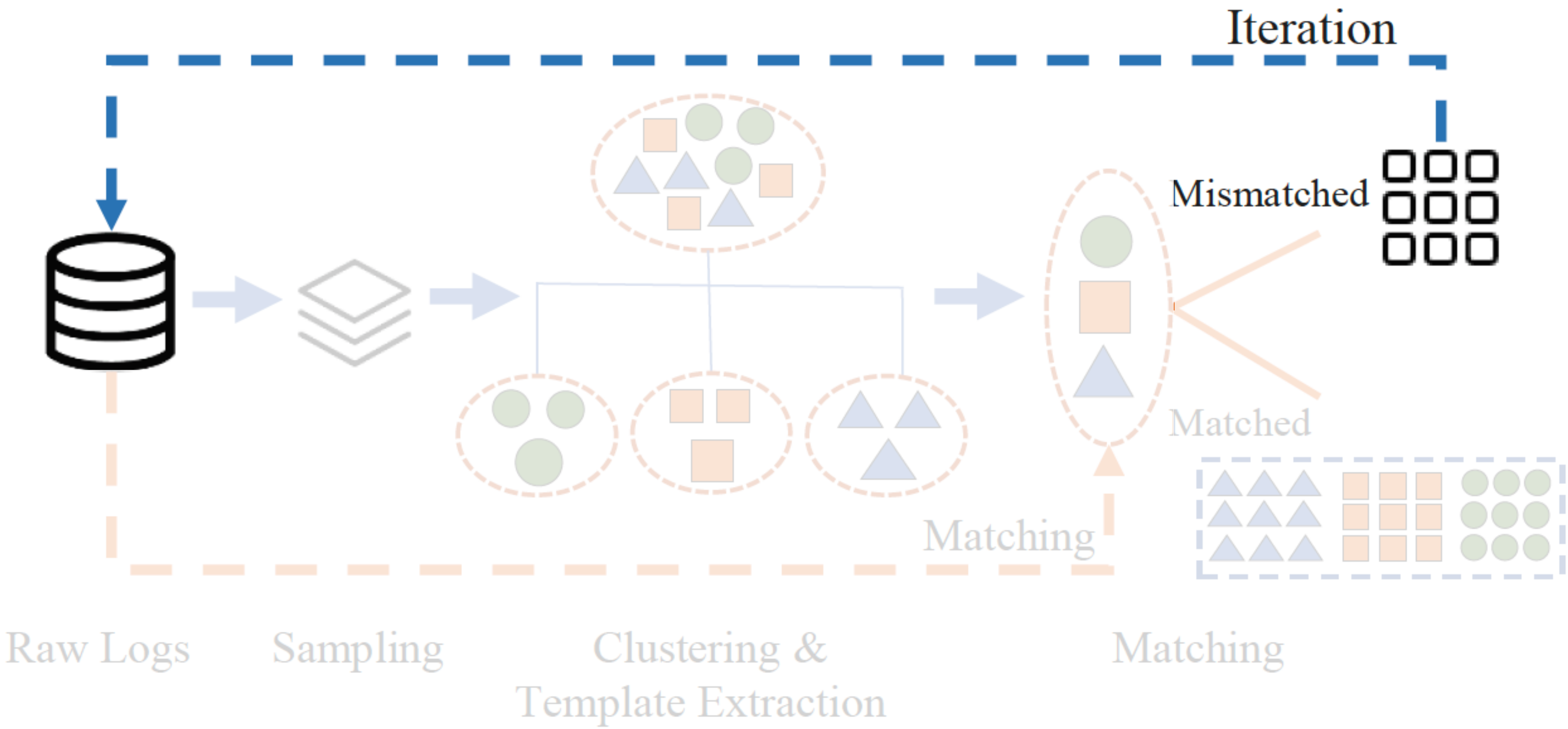
Build prefix tree

A B C *
A B G
A K E
A K F

Input 2: Templates

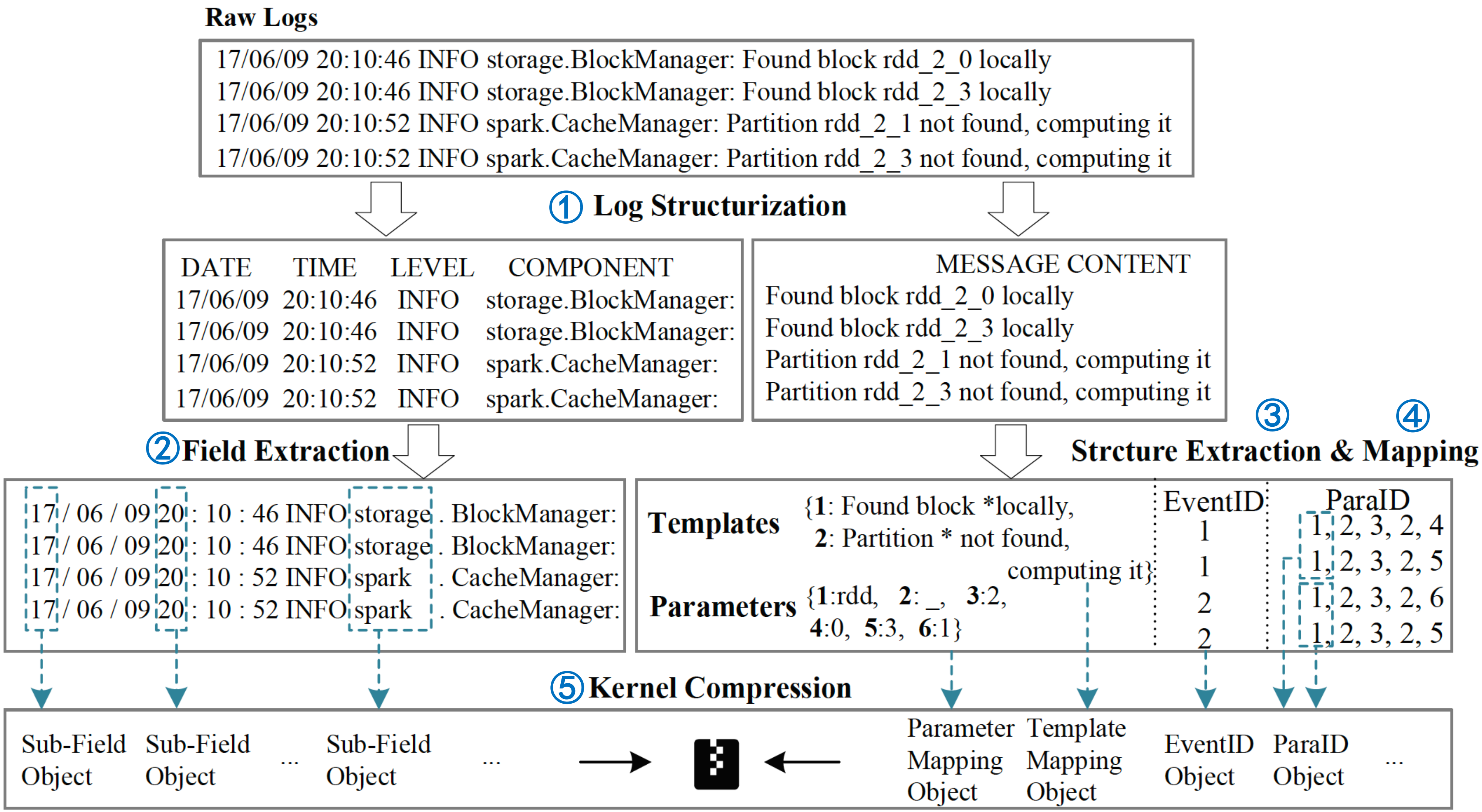
Content	Template ID	Template	Parameters
A B C D	E1	A B C *	D

Method: Iterative Structure Extraction (ISE)



Mismatched data goes through the process iteratively

Method: Example of Logzip



Raw Logs

```
17/06/09 20:10:46 INFO storage.BlockManager: Found block rdd_2_0 locally
17/06/09 20:10:46 INFO storage.BlockManager: Found block rdd_2_3 locally
17/06/09 20:10:52 INFO spark.CacheManager: Partition rdd_2_1 not found, computing it
17/06/09 20:10:52 INFO spark.CacheManager: Partition rdd_2_3 not found, computing it
```

① Log Structurization

DATE	TIME	LEVEL	COMPONENT
17/06/09	20:10:46	INFO	storage.BlockManager:
17/06/09	20:10:46	INFO	storage.BlockManager:
17/06/09	20:10:52	INFO	spark.CacheManager:
17/06/09	20:10:52	INFO	spark.CacheManager:

MESSAGE CONTENT
Found block rdd_2_0 locally
Found block rdd_2_3 locally
Partition rdd_2_1 not found, computing it
Partition rdd_2_3 not found, computing it

②Field Extraction

```
17 / 06 / 09 20 : 10 : 46 INFO storage . BlockManager:
17 / 06 / 09 20 : 10 : 46 INFO storage . BlockManager:
17 / 06 / 09 20 : 10 : 52 INFO spark   . CacheManager:
17 / 06 / 09 20 : 10 : 52 INFO spark   . CacheManager:
```

③ ④ Strcture Extraction & Mapping

Templates {**1**: Found block *locally, **2**: Partition * not found, computing it}

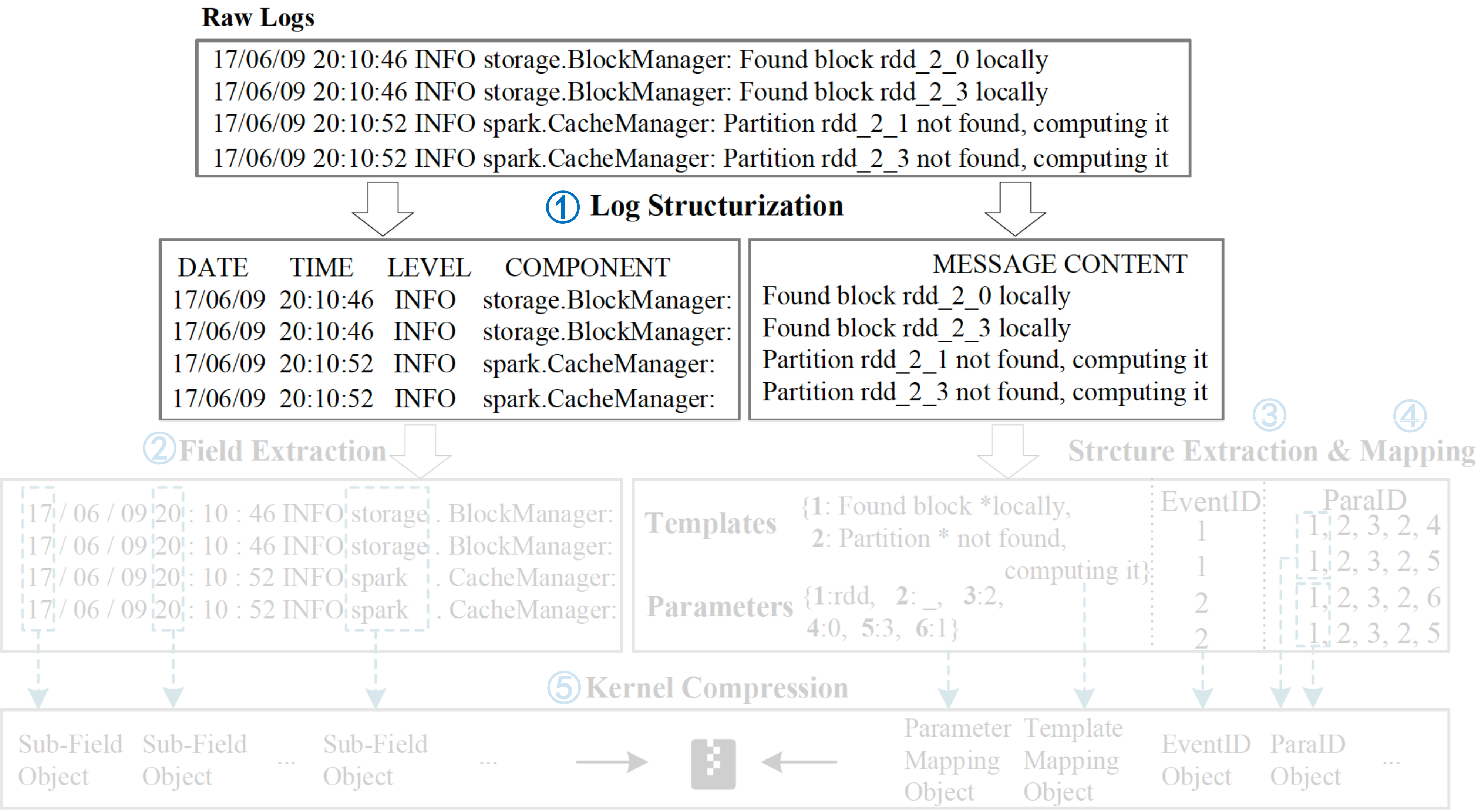
Parameters {**1**:rdd, **2**: _, **3**:2, **4**:0, **5**:3, **6**:1}

EventID	ParaID
1	1, 2, 3, 2, 4
1	1, 2, 3, 2, 5
2	1, 2, 3, 2, 6
2	1, 2, 3, 2, 5

⑤Kernel Compression

Sub-Field Object   Sub-Field Object   …   Sub-Field Object   …   →   ←   Parameter Mapping Object   Template Mapping Object   EventID Object   ParaID Object   …

Method: Example of Logzip



Raw Logs

```
17/06/09 20:10:46 INFO storage.BlockManager: Found block rdd_2_0 locally
17/06/09 20:10:46 INFO storage.BlockManager: Found block rdd_2_3 locally
17/06/09 20:10:52 INFO spark.CacheManager: Partition rdd_2_1 not found, computing it
17/06/09 20:10:52 INFO spark.CacheManager: Partition rdd_2_3 not found, computing it
```

① Log Structurization

DATE	TIME	LEVEL	COMPONENT
17/06/09	20:10:46	INFO	storage.BlockManager:
17/06/09	20:10:46	INFO	storage.BlockManager:
17/06/09	20:10:52	INFO	spark.CacheManager:
17/06/09	20:10:52	INFO	spark.CacheManager:

MESSAGE CONTENT
Found block rdd_2_0 locally
Found block rdd_2_3 locally
Partition rdd_2_1 not found, computing it
Partition rdd_2_3 not found, computing it

②Field Extraction

```
17 / 06 / 09 20 : 10 : 46 INFO storage . BlockManager:
17 / 06 / 09 20 : 10 : 46 INFO storage . BlockManager:
17 / 06 / 09 20 : 10 : 52 INFO spark   . CacheManager:
17 / 06 / 09 20 : 10 : 52 INFO spark   . CacheManager:
```

③ ④ Strcture Extraction & Mapping

Templates {**1**: Found block *locally, **2**: Partition * not found, computing it}

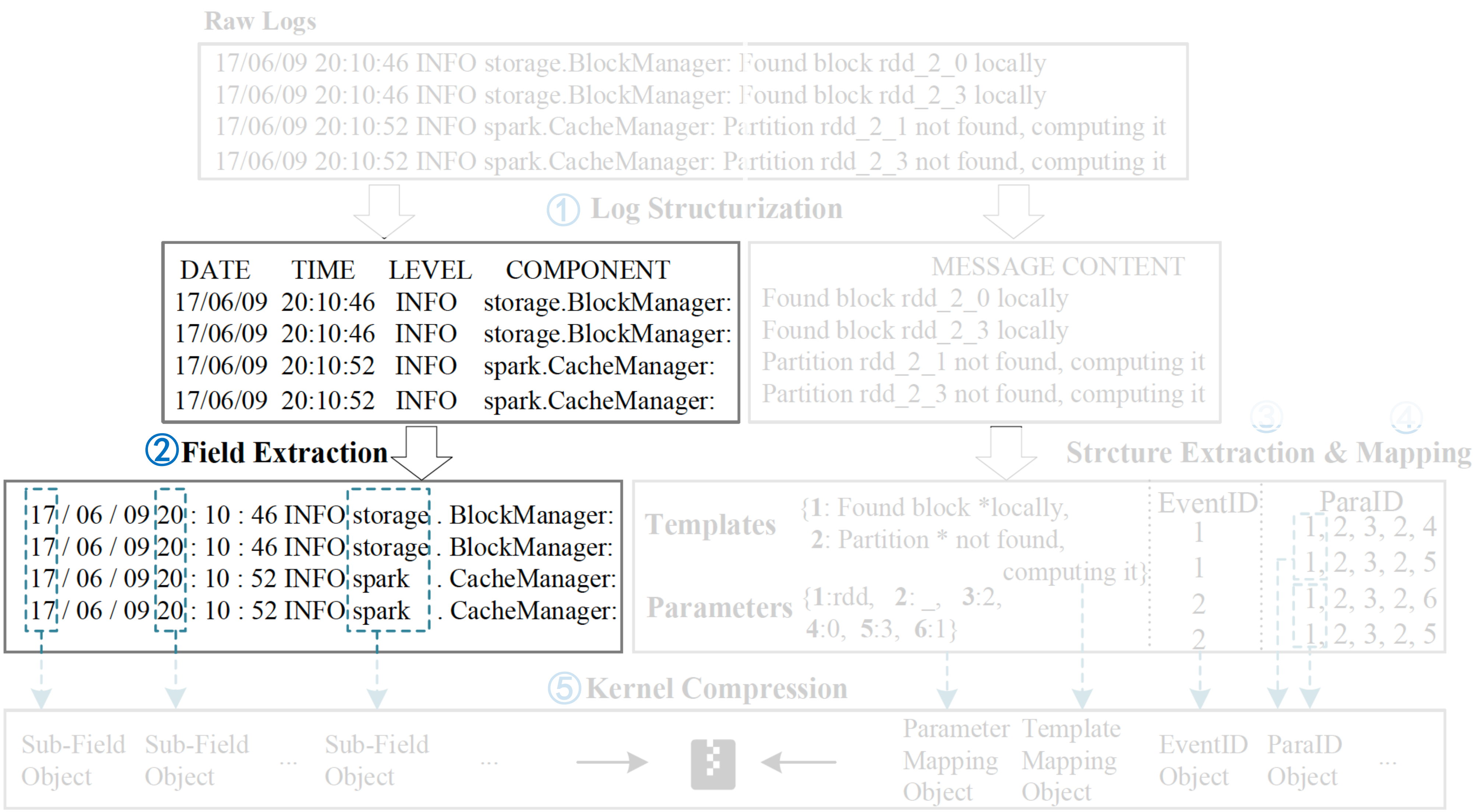
Parameters {**1**:rdd, **2**: _, **3**:2, **4**:0, **5**:3, **6**:1}

EventID	ParaID
1	1, 2, 3, 2, 4
1	1, 2, 3, 2, 5
2	1, 2, 3, 2, 6
2	1, 2, 3, 2, 5

⑤Kernel Compression

Sub-Field Object Sub-Field Object ... Sub-Field Object ... → ← Parameter Mapping Object Template Mapping Object EventID Object ParaID Object ...

Method: Example of Logzip



Raw Logs

```
17/06/09 20:10:46 INFO storage.BlockManager: Found block rdd_2_0 locally
17/06/09 20:10:46 INFO storage.BlockManager: Found block rdd_2_3 locally
17/06/09 20:10:52 INFO spark.CacheManager: Partition rdd_2_1 not found, computing it
17/06/09 20:10:52 INFO spark.CacheManager: Partition rdd_2_3 not found, computing it
```

① Log Structurization

DATE	TIME	LEVEL	COMPONENT
17/06/09	20:10:46	INFO	storage.BlockManager:
17/06/09	20:10:46	INFO	storage.BlockManager:
17/06/09	20:10:52	INFO	spark.CacheManager:
17/06/09	20:10:52	INFO	spark.CacheManager:

MESSAGE CONTENT
Found block rdd_2_0 locally
Found block rdd_2_3 locally
Partition rdd_2_1 not found, computing it
Partition rdd_2_3 not found, computing it

②Field Extraction

```
17 / 06 / 09 20 : 10 : 46 INFO storage . BlockManager:
17 / 06 / 09 20 : 10 : 46 INFO storage . BlockManager:
17 / 06 / 09 20 : 10 : 52 INFO spark   . CacheManager:
17 / 06 / 09 20 : 10 : 52 INFO spark   . CacheManager:
```

③ ④ Strcture Extraction & Mapping

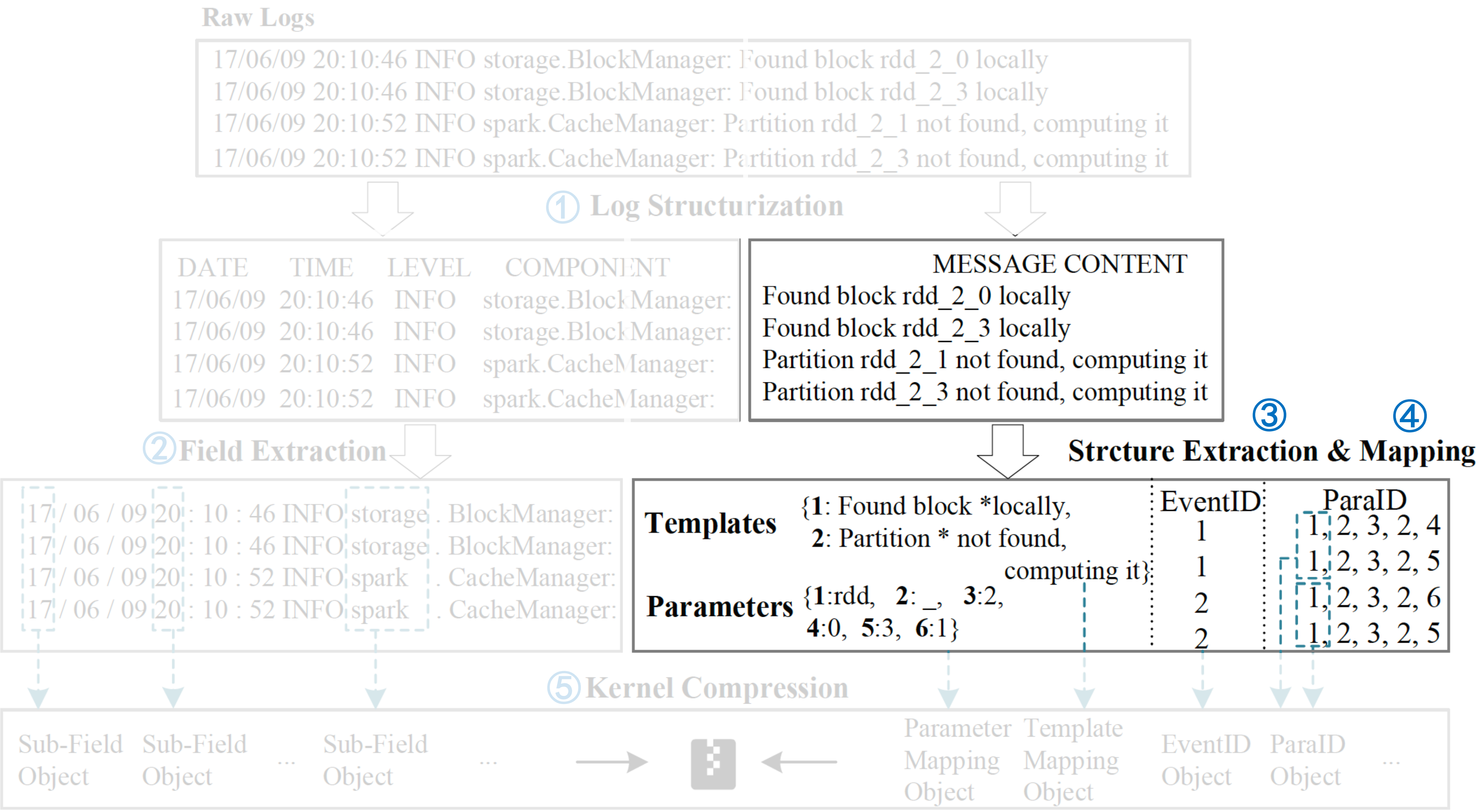
Templates {1: Found block *locally, 2: Partition * not found, computing it}

Parameters {1:rdd, 2: _, 3:2, 4:0, 5:3, 6:1}

EventID	ParaID
1	1, 2, 3, 2, 4
1	1, 2, 3, 2, 5
2	1, 2, 3, 2, 6
2	1, 2, 3, 2, 5

⑤Kernel Compression

Method: Example of Logzip



Raw Logs

```
17/06/09 20:10:46 INFO storage.BlockManager: Found block rdd_2_0 locally
17/06/09 20:10:46 INFO storage.BlockManager: Found block rdd_2_3 locally
17/06/09 20:10:52 INFO spark.CacheManager: Partition rdd_2_1 not found, computing it
17/06/09 20:10:52 INFO spark.CacheManager: Partition rdd_2_3 not found, computing it
```

① Log Structurization

DATE	TIME	LEVEL	COMPONENT
17/06/09	20:10:46	INFO	storage.BlockManager:
17/06/09	20:10:46	INFO	storage.BlockManager:
17/06/09	20:10:52	INFO	spark.CacheManager:
17/06/09	20:10:52	INFO	spark.CacheManager:

MESSAGE CONTENT

```
Found block rdd_2_0 locally
Found block rdd_2_3 locally
Partition rdd_2_1 not found, computing it
Partition rdd_2_3 not found, computing it
```

②Field Extraction

```
17 / 06 / 09 20 : 10 : 46 INFO storage . BlockManager:
17 / 06 / 09 20 : 10 : 46 INFO storage . BlockManager:
17 / 06 / 09 20 : 10 : 52 INFO spark   . CacheManager:
17 / 06 / 09 20 : 10 : 52 INFO spark   . CacheManager:
```

③ ④ Strcture Extraction & Mapping

Templates {**1**: Found block *locally, **2**: Partition * not found, computing it}

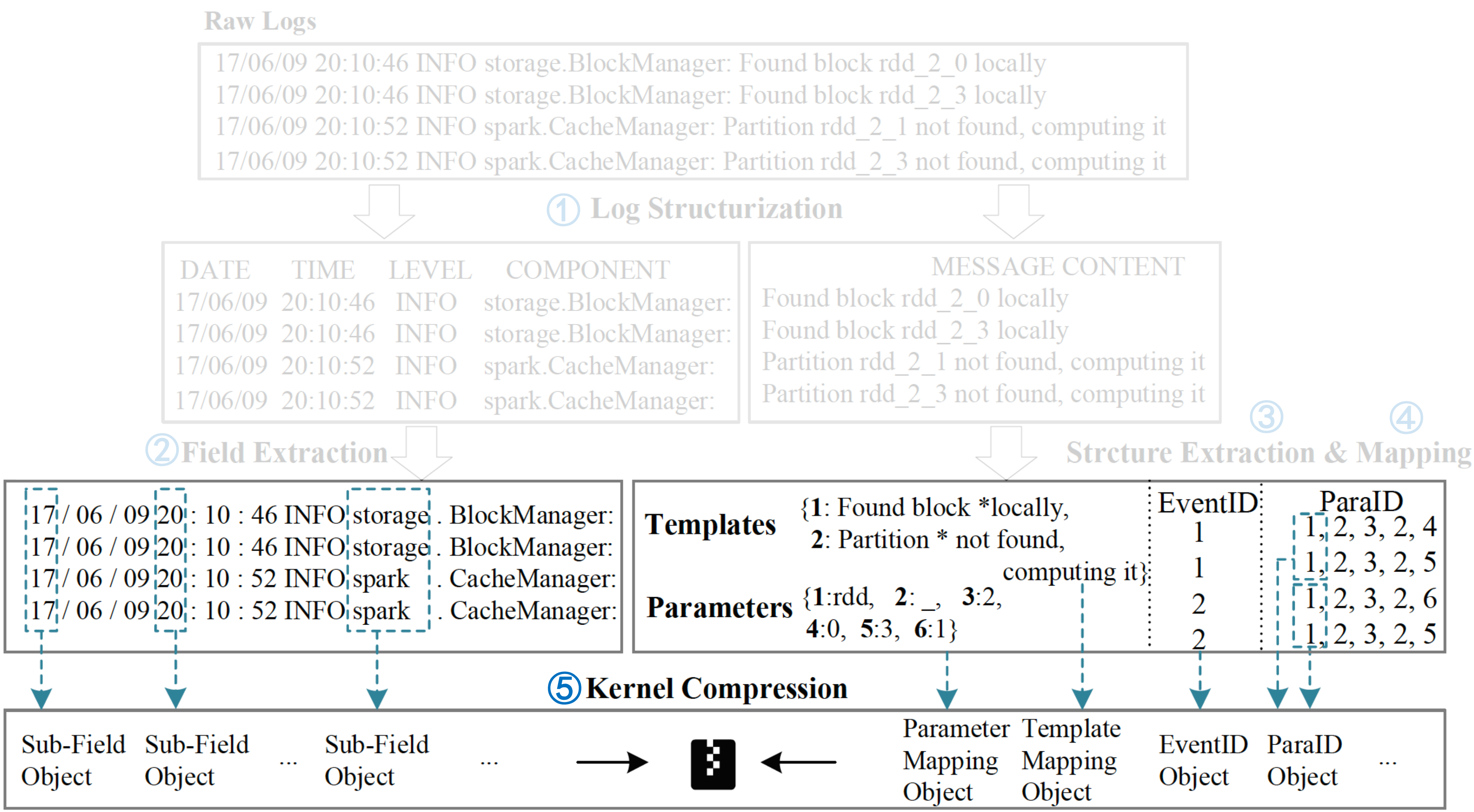
Parameters {**1**:rdd, **2**: _, **3**:2, **4**:0, **5**:3, **6**:1}

EventID	ParaID
1	1, 2, 3, 2, 4
1	1, 2, 3, 2, 5
2	1, 2, 3, 2, 6
2	1, 2, 3, 2, 5

⑤Kernel Compression

Sub-Field Object, Sub-Field Object ... Sub-Field Object ... → ← Parameter Mapping Object, Template Mapping Object, EventID Object, ParaID Object ...

Method: Example of Logzip



Raw Logs

```
17/06/09 20:10:46 INFO storage.BlockManager: Found block rdd_2_0 locally
17/06/09 20:10:46 INFO storage.BlockManager: Found block rdd_2_3 locally
17/06/09 20:10:52 INFO spark.CacheManager: Partition rdd_2_1 not found, computing it
17/06/09 20:10:52 INFO spark.CacheManager: Partition rdd_2_3 not found, computing it
```

① Log Structurization

DATE	TIME	LEVEL	COMPONENT
17/06/09	20:10:46	INFO	storage.BlockManager:
17/06/09	20:10:46	INFO	storage.BlockManager:
17/06/09	20:10:52	INFO	spark.CacheManager:
17/06/09	20:10:52	INFO	spark.CacheManager:

MESSAGE CONTENT
Found block rdd_2_0 locally
Found block rdd_2_3 locally
Partition rdd_2_1 not found, computing it
Partition rdd_2_3 not found, computing it

②Field Extraction

```
17 / 06 / 09 20 : 10 : 46 INFO storage . BlockManager:
17 / 06 / 09 20 : 10 : 46 INFO storage . BlockManager:
17 / 06 / 09 20 : 10 : 52 INFO spark   . CacheManager:
17 / 06 / 09 20 : 10 : 52 INFO spark   . CacheManager:
```

③ ④ Strcture Extraction & Mapping

Templates {**1**: Found block *locally, **2**: Partition * not found, computing it}

Parameters {**1**:rdd, **2**: _, **3**:2, **4**:0, **5**:3, **6**:1}

EventID	ParaID
1	1, 2, 3, 2, 4
1	1, 2, 3, 2, 5
2	1, 2, 3, 2, 6
2	1, 2, 3, 2, 5

⑤Kernel Compression

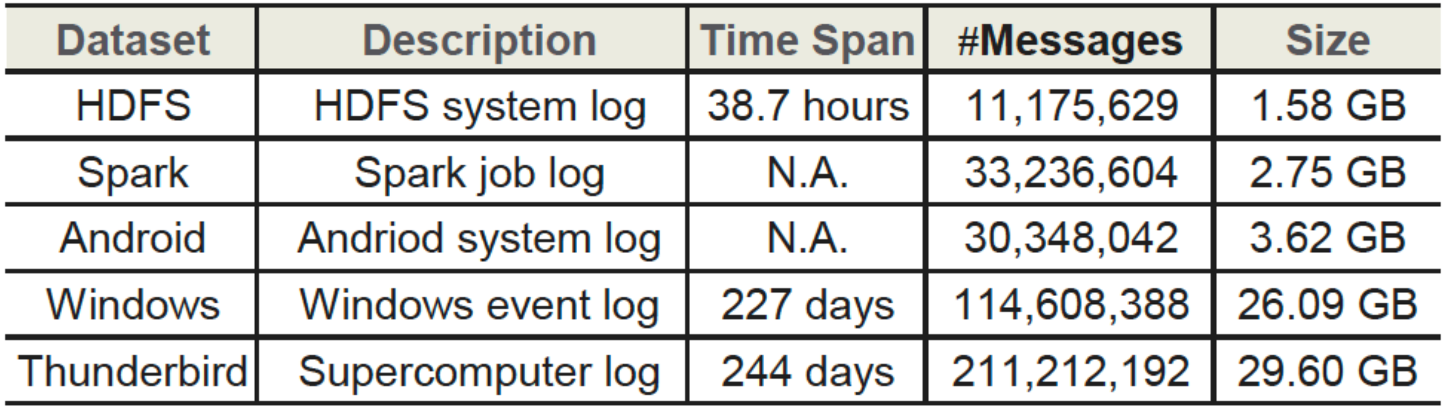
Sub-Field Object   Sub-Field Object   ...   Sub-Field Object   ...   →   ←   Parameter Mapping Object   Template Mapping Object   EventID Object   ParaID Object   ...

Experiment: Setup

Evaluation Metric:

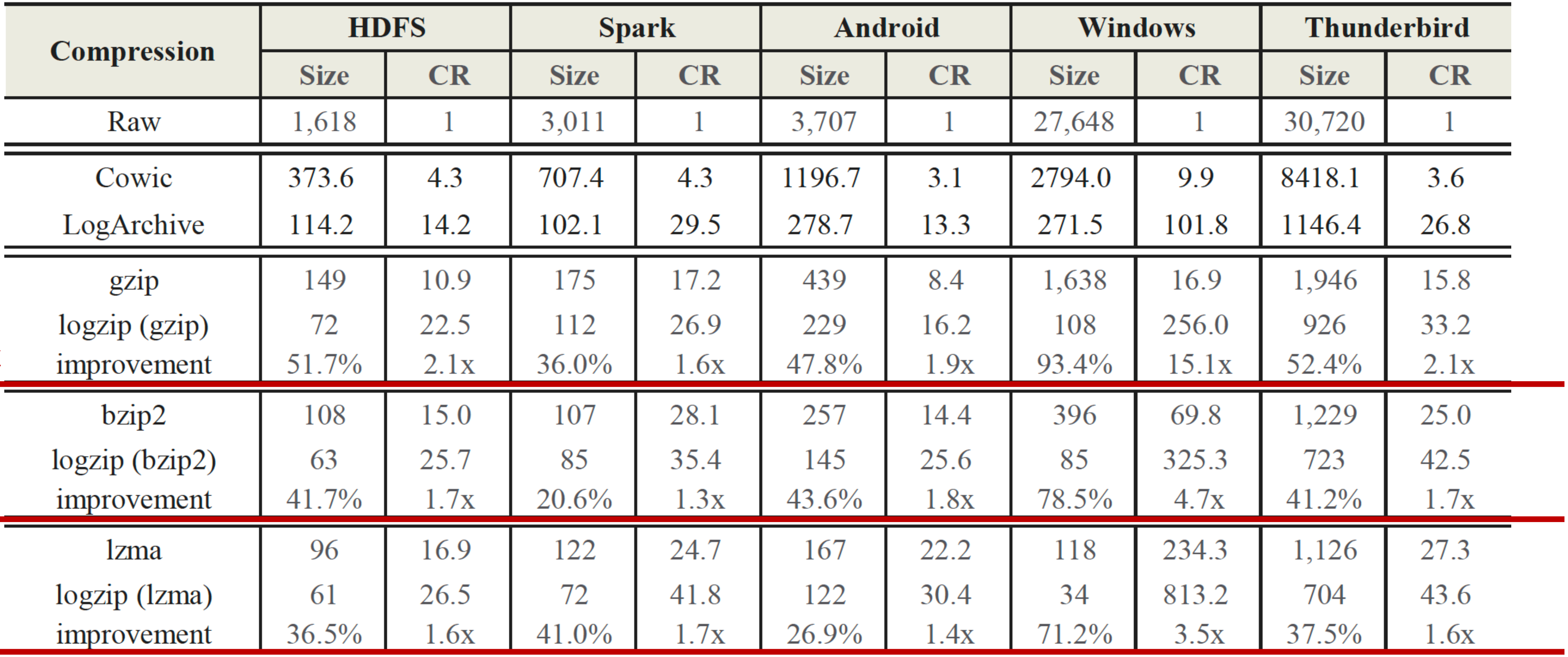
Compression Ratio (CR) $CR = \frac{Original\ File\ Size}{Compressed\ File\ Size}$ (The larger the better)

Dataset:



	Dataset	Description	Time Span	#Messages	Size
Distributed System	HDFS	HDFS system log	38.7 hours	11,175,629	1.58 GB
	Spark	Spark job log	N.A.	33,236,604	2.75 GB
Mobile System	Android	Andriod system log	N.A.	30,348,042	3.62 GB
Operating System	Windows	Windows event log	227 days	114,608,388	26.09 GB
Supercomputer	Thunderbird	Supercomputer log	244 days	211,212,192	29.60 GB

Experiment: Effectiveness of Logzip (level 3)



Compression	HDFS		Spark		Android		Windows		Thunderbird	
	Size	CR	Size	CR	Size	CR	Size	CR	Size	CR
Raw	1,618	1	3,011	1	3,707	1	27,648	1	30,720	1
Cowic	373.6	4.3	707.4	4.3	1196.7	3.1	2794.0	9.9	8418.1	3.6
LogArchive	114.2	14.2	102.1	29.5	278.7	13.3	271.5	101.8	1146.4	26.8
gzip	149	10.9	175	17.2	439	8.4	1,638	16.9	1,946	15.8
logzip (gzip)	72	22.5	112	26.9	229	16.2	108	256.0	926	33.2
improvement	51.7%	2.1x	36.0%	1.6x	47.8%	1.9x	93.4%	15.1x	52.4%	2.1x
bzip2	108	15.0	107	28.1	257	14.4	396	69.8	1,229	25.0
logzip (bzip2)	63	25.7	85	35.4	145	25.6	85	325.3	723	42.5
improvement	41.7%	1.7x	20.6%	1.3x	43.6%	1.8x	78.5%	4.7x	41.2%	1.7x
lzma	96	16.9	122	24.7	167	22.2	118	234.3	1,126	27.3
logzip (lzma)	61	26.5	72	41.8	122	30.4	34	813.2	704	43.6
improvement	36.5%	1.6x	41.0%	1.7x	26.9%	1.4x	71.2%	3.5x	37.5%	1.6x

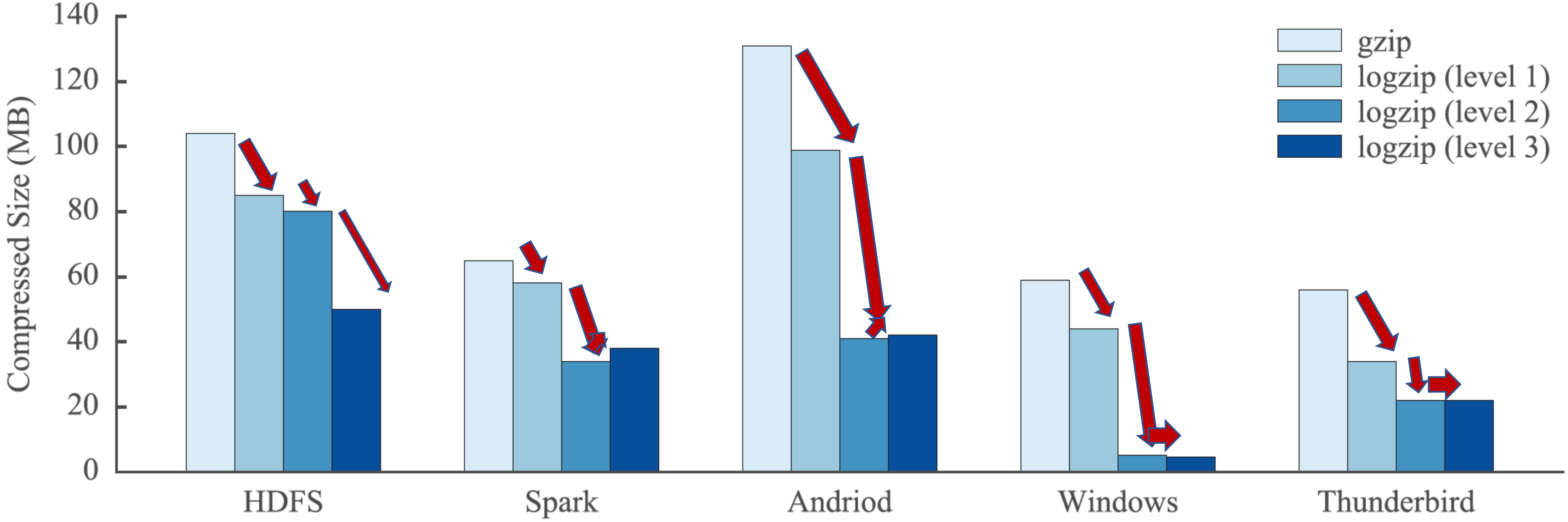
gzip
avg: ~2.0 x
max: 15.1x

bzip2
avg: ~1.6x
max: 4.7x

lzma
avg: ~1.6x
max: 3.5x

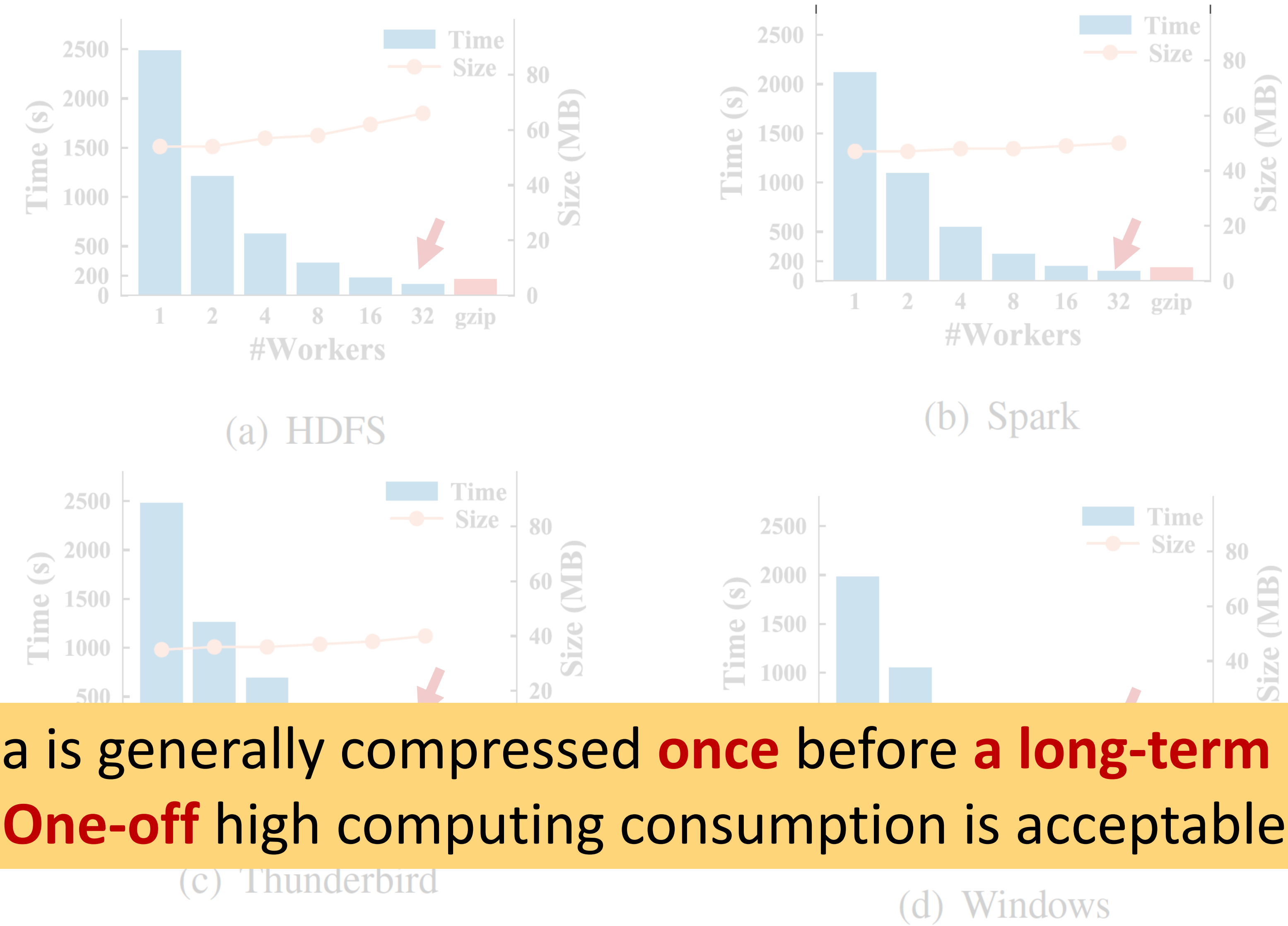
Experiment: Effectiveness of Logzip in each level

Level 3: Structure extraction, with parameter mapping



CR logzip (gzip) achieved by processing first 1GB data of each datasets

Experiment: Efficiency of Logzip (Level 3)



(a) HDFS

(b) Spark

Log data is generally compressed **once** before **a long-term** storage

One-off high computing consumption is acceptable

(c) Thunderbird

(d) Windows

Compression Time & Size vs #Worker

Conclusion: Logzip

- ✓ Iterative clustering for **hidden structure extraction**
- ✓ Efficient template **matching**
- ✓ Effective and efficient log data **compression**

LOGPAI

Log Analysis Powered by AI

www.logpai.com

LogAdvisor

- Learning to log: A framework for determining optimal logging points

[ICSE’14, ICSE’15]

Loglizer

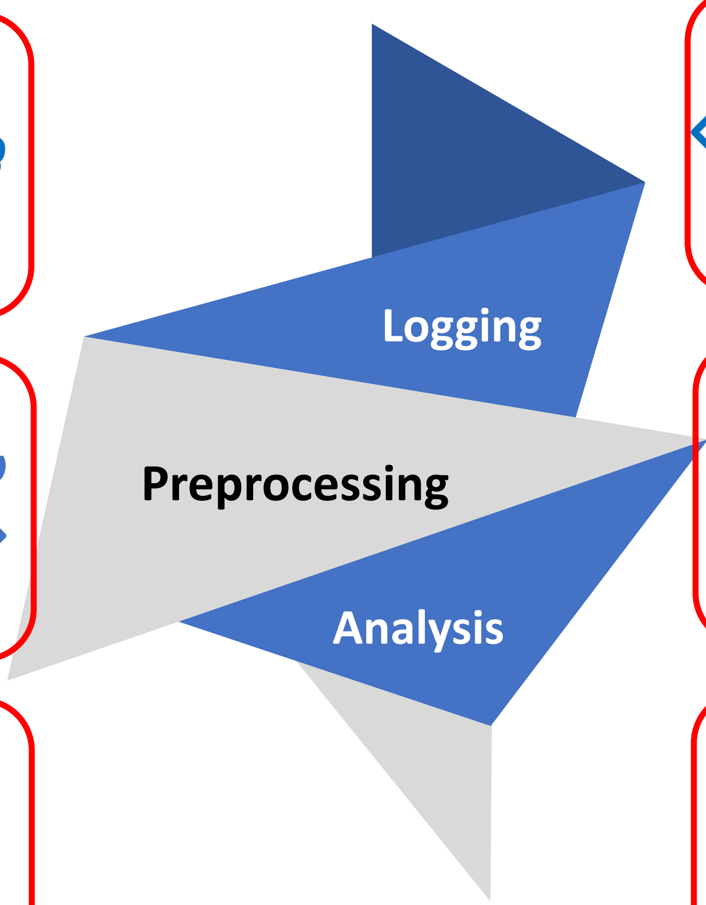
- A log analysis toolkit for automated anomaly detection

[ISSRE'16]

LogParser

- A toolkit for automated log parsing

[ICSE'19, TDSC'18, DSN'16]



LoggingDescriptions

- A collection of Software Logging Statements

[ASE’18]

Log3C

- Log-based Problem Identification

[FSE’18]

Logzip

- An effective and efficient log compression tool

[ASE’19]

LOGPAI

Questions and Cooperation are welcome!

🔗 www.logpai.com

✉ info@logpai.com